AD-A233 041

:UMENTATION PAGE

Form Approved
OMB No. 0704-0188

tion is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, ...ething and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this ...ducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson ..., and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
22 January 1991	Final Technical, 30 Sept., 1986-31 Dec., 1989

4. TITLE AND SUBTITLE

Auditory-Acoustic Basis of Consonant Perception - Attch A-I

5. FUNDING NUMBERS

G - AFOSR-86-0335
PE - 61102F
PR - 2313
TA - A6

(2)

6. AUTHOR(S)

James D. Miller

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Central Institute for the Deaf
818 S. Euclid Avenue
St. Louis, MO 63110

8. PERFORMING ORGANIZATION REPORT NUMBER

AFOSR-TR- 91 0130

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research/NL
Bolling Air Force Base Bld 410
District of Columbia 20332

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

DTIC COPY

11. SUPPLEMENTARY NOTES

DTIC
ELECTE
MAR 08 1991
S D
D

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified/Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

New facts of this auditory-acoustic basis of consonant perception were discovered as listed. 1) Plosive consonants can be distinguished from fricative consonants by the peak rate of rise of intensity at their onsets. 2) The acoustic characteristics that serve to identify plosive bursts and voiceless fricatives by place of articulation can be usefully described in terms of formants defined by a novel algorithm. 3) A connectionist software model that examines fourteen psychophysically relevent acoustic measures can classify any acoustic segment of speech by the location of its source in the talker's vocal tract. 4) Preliminary studies of sonorant and nasal consonants have identified the putative acoustic cues for their identification by human listeners and/or machines. 5) New methods for formant tracking were developed. 6) An important set of software tools were developed that allow further studies of the auditory-acoustic basis of consonant perception. 7) These tools have also aided in the studies of vowels and diphthongs, whose characteristics are being elucidated under primary support from the National Institutes of Health.

14. SUBJECT TERMS

15. NUMBER OF PAGES 20

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	PR	- Project
G	- Grant	TA	- Task
PE	- Program Element	WU	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- DOE - See authorities.
- NASA - See Handbook NHB 2200.2.
- NTIS - Leave blank.

Block 12b. Distribution Code.

- DOD - Leave blank.
- DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- NASA - Leave blank.
- NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

AUDITORY-ACOUSTIC BASIS OF CONSONANT PERCEPTION

James D. Miller
Central Institute for the Deaf
818 S. Euclid Avenue
St. Louis, MO 63110

22 January 1991

Final Technical Report for Period 30 September 1986 - 31 December 1989

Prepared for

Department of the Air Force
Air Force Office of Scientific Research (AFSC)
Bolling Air Force Base, DC 20332-6448

Accesion For
NTIS CRA&I
DTIC TAB
Unannounced
Justification
By
Distribution /
Availability Codes
Dist | Avail and/or Special
A-1
QUALITY INSPECTED 3

91 3 06 136

Final Technical Report
Grant AFOSR 86-0335
For the period 30 September 1986 - 31 December 1989

I. SUMMARY

In the three years of this grant, several projects were undertaken. All projects aimed at examining the acoustic characteristics of American English consonants and developing algorithms that would uniquely and correctly discriminate consonants, either as groups (e.g., stops vs. fricatives) or as single phonetic units (e.g., bilabial vs. alveolar stop).

In the process of carrying out these projects large data sets of natural speech, produced by native speakers of American English, were collected, analyzed, and evaluated. On the basis of these results two algorithms were developed, one for characterizing the spectra of voiceless stops and fricatives and one for discriminating between stops and fricatives, regardless of voicing. In another project, connectionist software system was developed that classifies each acoustic segment (12 msec) of speech by source in the talker's vocal tract and this is essential for the appropriate spectral description of speech waveforms. Methods of automatic formant tracking were developed that perform well for certain sets of syllables. It is believed that more effort could generalize these methods to running speech. Additionally, preliminary descriptions of the defining acoustic characteristics of stop consonants (p,t,k,b,d,g), of voiceless fricatives (s,sh,f,th,h), liquids and glides (w,l,r,j) and nasals (m,n,ng) were developed. Furthermore, in conjunction with work done under Grant NIDCD-5 R01 DC00296-06, several software packages were developed for data visualization within the model of the auditory-perceptual theory of speech perception proposed by Miller (1989) (section III.D.4). Parallel studies of vowels and diphthongs supported primarily by the NIH advanced our knowledge of their acoustic correlates and has led to the notion of auditory-acoustic target zones for steady-state and for spectral glides.

II. RESEARCH OBJECTIVES

The overall objective of the research under this grant, and under Grant NIDCD -5 R01 DC00296-06, is to provide a detailed description of the process by which a listener decodes the incoming acoustic signal and generates an internal phonetic representation of the speaker's utterance. This description, within the auditory perceptual theory of phonetic recognition, should provide the basis for automatic, speaker-independent recognition of continuous speech.

To this end, naturally produced utterances of the consonant sounds of English, divided into classes according to their source characteristics, were subjected to various analyses using computer-implemented techniques. In addition, a significant effort has been made to refine and further develop the visual aids which illustrate the hypothesized structure within the three-dimensional space of the auditory-perceptual theory.

III. STATUS OF THE RESEARCH

Introduction

The goal of this research project was to develop a detailed and comprehensive account of exactly how the human listener decodes the acoustic speech signal into a string of phonetic elements that then provide a basis for the recognition of the words of a spoken language. The eventual goal was to arrive at a description of this process in a manner so detailed that computer programs could exactly simulate this process and serve then as a phonetic typewriter. The original proposal to the AFOSR presented a comprehensive theory describing this process and outlined an extensive five year plan to accomplish the essentials of these objectives. The plan was conservatively developed and an average annual budget of $320,000 was requested. The AFOSR approved our proposal but annual funding was sharply reduced to about 40% of the requested amount and the time period of funding also was reduced from 5 years to 3 years. Furthermore, a requested extension after completion of the first three years of work severely damaged the success of the project as it was necessary to eliminate personnel thus making it impossible to fully capitalize on the important gains made during the first three years.

In spite of these problems, very significant advances were made during the three-year grant period. In sum, all of our research was consistent with our initial projection that a conceptual model of the processes whereby the human listener converts the acoustic signal into a string of phonetic elements could be successfully implemented. The actualization of such a system only requires a few more years of adequate effort.

The main achievements of the three-year grant period were as follows:

1) An algorithm was developed which allows automatic description of burst-friction sounds, that is, the bursts of plosive consonants, the aspirations of voiceless plosive consonants, and the voiced and voiceless fricatives in terms of formant center frequencies in a manner consistent with the description of formants of vocalic sounds. A description of this algorithm is in a manuscript (Jongman and Miller - section III.A.2 of this report) and its application has been described in two presentations at scientific meetings - [Miller and Jongman (1987) section III.B.4 and Jongman, A. (1987) "Locations of burst onsets of stops in the auditory-perceptual space." J. Acoust. Soc. Am. 82(S1), S82(A)]. This alogrithm marks the first systematic attempt at a simple description of burst-friction sounds in terms that are directly comparable to the description of vocalic sounds. Although experience with this algorithm suggests bases for its improvement, it has served very well to allow classification of the phonetic correlates of stop-bursts and of fricatives.

2) Preliminary target zones for voiceless fricatives and the bursts of voiceless stops. Using the algorithm described above, Weigelt and Miller conducted numerous preliminary analyses of the bursts of voiceless stops (P,T,K) and the spectra of voiceless fricatives (F,TH,S,SH,HH). Preliminary spectral target zones performed well in classifying these tokens.

3) An algorithm was developed that correctly distinguishes plosive consonants from fricative consonants. This work is described in a publication (III.A.3) and in a manuscript submitted for publication (III.A.4). This algorithm finds the peak rate-of-rise in intensity at the onset of the consonant

and uses this to classify the consonant as a plosive or a fricative. Performance level of the algorithm is about 96% correct.

4) Essential to the correct analysis of speech is the ability to classify each acoustic segment of speech by the location of the sound source in the talker's vocal track. Each segment of speech falls into one of four categories: No source (silence), glottal source, supraglottal source, or both glottal and supraglottal. These four categories are referred to as Silence, Glottal Source, Burst-Friction, and Mixed. In a completed dissertation (III.A.5), Sadoff described a connectionist software system that correctly classified the segments in continuous speech with about 90% accuracy. This accomplishment marks an important forward step in the progress toward automatic speech analysis and recognition.

5) Preliminary target zones were developed for the nasal consonants [M,N,NX] and for the glides and liquids [U,W,R,Y]. Based on mesurements of 48 tokens produced by two male and two female talkers of the nasals following four different vowels, it was possible to develop target zones in the auditory-perceptual space for the steady-state portions of the nasal consonants. Based in measurements of 96 tokens produced by one male and one female talker preliminary target zone locations were developed for L,W,R,&Y. In each case these target zones overlap to some degree with those of other phonetic elements and it appears that additional criteria need to be used to distinguish these. The criteria suggested by these preliminary studies appear to be rather easily defined. To distinguish L's from the vowel [UH] and the consonant [W], which sometimes have similar spectral shapes, one notes the deep spectral valley around 2200 Hz and the associated narrowing of the bandwidth of F3 in L's as opposed to UH's and W's. W's which are often spectrally similar to the vowels [UW] and [UH] and with the consonant [L], exhibit characteristic patterns of formant change (spectral glides) which serve to distinguish them from the vowel sounds and do not have the spectral dip and narrowing of F3 characteristic of L's. The consonant [Y] is often similar in spectral shape to the vowels [IY] and [IH]. However, it can be distinguished from these by characteristic patterns of formant change that result in a "loop" in the spectral path as it travels through the auditory-perceptual space. The consonant R appears to be spectrally nearly identical to the vowel ER. The distinction sometimes appears to be based on factors such as position within the syllable and duration and other times appears to be irrelevent. Some of these results have been reported at a scientific meeting (Miller, 1988 — section III.B.1 of this report).

6) Important initiatives were made in a) automatic formant tracking (III.B.3) and b) in the calculation of the loudness of speech and the effects of spectral goodness and loudness in determining the perceived characteristics of speech sounds (III.B.2).

7) A whole set of software necessary for our approach to the study of speech was developed. For descriptions, see section III.D of this report.

8) Support from AFOSR assisted in the completion of several studies of vowels and diphthongs — see Sections III.E, III.F of this report.

In summary, under AFOSR support we were able to develop methods and gather data that gave validity to the PI's auditory-perceptual approach to human speech

perception and automatic speech recognition. Only limitations of time and funding seem to have slowed the attainment of the objectives of the original proposal.

Organization of the Report

This section is organized in sections pertaining to research conducted under funding from this grant, under joint funding from NIDCD-5 R01 DC00296-06, and research in general by the participants. The sections also are divided into projects completed (published or submitted), projects reported at scientific meetings, and projects in progress where applicable.

A. Projects completed and published, accepted for publication, or submitted. (AFOSR Funding)

1. Chang, H.M. (1987). "SWIS: See what I say, a speaker-independent word recognition system by phoneme-oriented mapping on a phonetically-encoded auditory-perceptual speech map," Ph.D. dissertation, Washington University, St. Louis, MO.

   This paper reports the development of a prototype speaker-independent word recognition system based on a new model of phoneme perception proposed in the auditory-perceptual theory of phonetic recognition (see Appendix A). The system is able to map directly acoustical parameters to phonetic events by means of a phonetically-encoded auditory-perceptual map designed for a selected set of phonemes derived from all the stressed syllables in an 11-word digit vocabulary recorded from 21 new talkers (11 males and 10 females) for the digit vocabulary, a 61.5% score for phoneme recognition is obtained while the phoneme insertion rate is 4.5%. A 53% score is obtained for word recognition by using only the phonetic information derived from the vocalic segments of a stressed syllable in a word utterance.

   After reviewing the sources of success and failure it is argued that a phoneme-oriented system based on the broad framework established in the theory and implementated in this project, could be generalized to provide a solution to speaker-independent isolated word recognition for a medium-sized vocabulary of 50 to 200 words.

2. Jongman, A. and Miller, J.D. "Method for the location of burst-onset spectra in the auditory-perceptual space: A study of place of articulation in voiceless stop consonants," J. Acoust. Soc. Am. — In press.

   A method for distinguishing burst onsets of voiceless stop consonants in terms of place of articulation is described. Four speakers produced the voiceless stops in word-initial position in six vowel contexts. A metric was devised to extract the characteristic burst-friction components at burst onset. The burst-friction components, derived from the metric as sensory formants, were then transformed into log frequency ratios and plotted as points in an auditory-perceptual space (APS). In the APS, each place of articulation was seen to be associated with a distinct region, or

target zone. The metric was then applied to a test set of words with voiceless stops preceding ten different vowel contexts as produced by eight new speakers. The present method of analyzing voiceless stops in English enabled us to distinguish place of articulation in these new stimuli with 70% accuracy.

3. Weigelt, L., Sadoff, S.J., and Miller, J.D. (1990). "Plosive/fricative distinction: the voiceless case," J. Acoust. Soc. Am., 87(6), 2729-2737.

   Using only three measures of the waveform, the zero-crossing rate, the logarithm of the root-mean-square (rms) energy, and the derivative of the log rms energy with respect to time [termed rate of rise (ROR)], voiceless plosives (including affricates) can be distinguished from voiceless fricatives in word-initial, medial, and final positions. Peaks in the ROR contour are considered for signficance to the plosive/fricative distinction by examining the log rms energy and zero-crossing rate. Then, the magnitude of the first significant peak in the ROR contour is used as the primary classifier. The algorithm was tested on 134 tokens (72 word-initial tokens produced by four female and four male speakers; 360 word-medial tokens produced by two males and two females; 320 word-final tokens produced by two males and two females). Data from two male and two female speakers (360 word-initial tokens) were used as a training set, and the remaining data were used as a test set. The overall rate of correct classification was 96.8%. Implications of this result are discussed.

4. Weigelt, L., Sadoff, S.J., and Miller, J.D. "Plosive/fricative distinction: the voiced case." Submitted to Speech Communication.

   We previously reported an algorithm which distinguishes voiceless plosives from voiceless fricatives with a success rate of 96.8% (J. Acoust. Soc. Am. 87(6), 2729-2737). A similar, but modified, algorithm also makes this distinction for the voiced case. Here results are presented on the modified algorithm. The input signal is high-pass filtered and two measures of the resulting waveform are used. One measure is the log rms energy. The other is the derivitives of log rms energy over time or rate of rise (ROR) of amplitude. Peaks in the ROR contour are considered for significance to the plosive/fricative distinction by examining the log rms energy. Then, the magnitude of the first signficant peak in the ROR contour is used as the primary classifier. The resulting algorithm was tested on 1128 tokens (with the consonant of interest in word-initial, medial or final position) recorded in an anechoic chamber with 564 tokens serving as a training set and the remaining data serving as a test set. The overall success rate was 96.3%. This new algorithm, developed for the voiced case, was also applied to the previously studied voiceless data. The rate of correct classification across all word positions from both the voiced and voiceless cases was 95.8%.

5. Sadoff, S.J. (1990). "Classifying speech into silence, glottal source, burst friction, or mixed categories," Sc.D. dissertation, Washington University, St. Louis, MO.

   The primary goal of this research is to automatically segment speech into four acoustic categories based on the location of the sound sources in the

human vocal tract: silence (S), glottal source (G), burst friction (B), mixed (M). A multidisciplinary approach has been taken in an attempt to solve this interesting and important problem in speech processing. In addressing this problem, knowledge and techniques from many fields including, but not limited to digital signal processing, artificial intelligence, pattern recognition, neural networks, psychophysics, speech perception, speech production, and the acoustics of speech have been applied. A connectionist system has been implemented and trained on 34 seconds of continuous speech from a female speaker. When tested on an additional 87 seconds of speech from the same speaker, the classification was 90.0 percent correct and when tested on 106 seconds of speech from a male speaker the accuracy was 88.8 percent.

B. Projects completed and reported at scientific meetings (other than those under "A") (AFOSR Funding)

1. Miller, J.D. (1988). "Auditory-perceptual analysis of selected syllables." J. Acoust. Soc. Am., 84, S154(A).

   Analyses of consonant-vowel syllables (CVs) in terms of the auditory-perceptual theory of phonetic recognition will be presented. Examples of CVs will include a voiceless stop, a voiceless fricative, a nasal, and an approximant paired with monophthongal vowels. Spectral analyses are used to locate the formant peaks and to track these during the course of the syllable, producing a sequence of spectra, one for each ms of waveform. Formant and F0 information from these sequences is then converted into sensory and perceptual paths in the theory's auditory-perceptual space. This space contains subspaces, called perceptual target zones. The activation of a zone results in the output of a phonetic code. While the exact conditions for this activation are not yet known, it appears that certain aspects of the behavior of a perceptual path in relation to the perceptual target zones can determine the phonetic transcription of a syllable. [Work supported by NINCDS and AFOSR.]

2. Miller, J.D., Sadoff, S.J., and Veksler, M.R. (1988). "Sensory-perceptual transformations in speech analysis." J. Acoust. Soc. Am., 83, S70(A).

   In Miller's auditory-perceptual theory of phonetic recognition, the sensory effects of sequences of glottal-source spectra, burst-friction spectra, and silences are integrated into a unitary perceptual response by the sensory-perceptual transformations. The mathematical and computational implementations of the hypothesized transformations will be described. One transformation is applied to spectral pattern or shape and is implemented formant by formant. For example, the center frequency of a perceptual formant is calculated by second-order difference equations from the center frequencies, loudness, and bandwidths of the corresponding sensory formants of glottal-source and burst-friction spectra, when either or both are present, and from similar values for a neutral vocal tract during silence. The loudness of the perceptual response, which is calculated as the integrated loudness of the burst-friction and glottal-source inputs, decays slowly after the cessation of sensory input and, in this way, an audible perceptual response can be maintained during brief

periods of silence. Examples of application of these concepts to a variety of speech sounds, such as stops and fricatives, will be illustrated. [Work supported by NINCDS and AFOSR.]

3. Fingerhut, J.A. and Miller, J.D. (1989). "Automatic correction of formant tracks," J. Acoust. Soc. Am., 86(S1), S124 (ASA meeting, Fall 1989).

   Since commercially available formant trackers are often inconsistent, much hand-editing is required to correct their outputs. A software package is currently being developed that may reduce this problem. Two female and two male speakers were recorded producing ten vowels in S-vowel-T context, each vowel twice. Boundaries between burst-friction, glottal-source, and silent segments are located using zero crossing and rms energy measurements. Pitch and formant-frequency values are extracted using the API and SGM commands of ILS. Our software corrects the pitch contour and calculates a sensory reference from it [J.D. Miller, J. Acoust. Soc. Am. 85, 2114-2134 (1989)]. Rules based on the relation between peak frequencies and this sensory reference are utilized to label the peaks as formants and the resulting formant tracks are then low-pass filtered. In this way, satisfactory tracking is obtained for all 80 syllables. [Work supported by AFOSR and NIDCD.]

4. Miller, J.D. and Jongman, A. (1987). "Auditory-perceptual approach to stop consonants." J. Acoust. Soc. Am., 82(S1), S82(A).

   Using natural tokens and the synthetic stimuli of Abramson and Lisker [6th Internat. Congr. Phonet. (1970)], the auditory-perceptual theory shall be applied to syllable-initial stop consonants. The burst-friction components of each token are analyzed for the locations of their second and third sensory formants, BF2 and BF3. These are located in the auditory-perceptual space (APS) by the formulas: $x = \log(BF3/BF2)$ and $y = \log(BF2/SR)$, where SR is the sensory reference. The glottal-source components are then analyzed for their sensory formants: SF1, SF2, and SF3. These are then located in APS by equations: $x = \log(SF3/SF2)$, $z = \log(SF1/SR)$, and $z = \log(SF2/SF1)$. In this way, a sensory path comprised of burst-friction points and glottal-source points is generated. Next, a sensory-perceptual transformation is applied to generate a unitary perceptual path. Since the sensory-perceptual transformation is reactive, the perceptual path overshoots the burst-friction onset of the token and enters a physically unrealizable octant of APS, wherein the perceptual target zones for the stops are located. Distinct target zones are estimated for the stops [p,t,k,d,b,g], for h-like aspiration, and for voice bars. [Supported by AFOSR and NINCDS.]

C. Software development. (Joint AFOSR & NIDCD Funding)

We have made considerable progress in developing software for the implementation of the theory on computers, using both the Evans and Sutherland three-dimensional graphics terminal and regular two-dimensional terminals. Below we report the work of the last two years. The first year sponsored by the NIH Grant (5 R01 DC00296-06) and the second year jointly supported by the NIH Grant and the AFOSR Grant that is the subject of this report.

The Evans and Sutherland PS300, a high speed, high resolution color graphics system, and its VAX-VMS host system are used to display, manipulate, and analyze objects in the three-dimensional auditory-perceptual space. In the majority of cases, software used in this research effort has been specially developed for these rather specialized applications. The programs MWVNET, DISPLAY and SLICER are the three most used application programs and are described below.

MWVNET is a PS300 function network that allows the user to examine an object and manipulate it in four different coordinate systems: world, model, part and view. The program implements keyboard commands for the choice of coordinate systems and rotary dial input for scaling, translation and rotation of the displayed objects. MWVNET also forms the framework for most of the other application programs written for speech perception studies on the PS300.

DISPLAY is an application program whose primary function is to provide a user interface for the display and manipulation of objects defined in PS300 code. It has facilities for highlighting, blinking, coloring and hiding objects. It also provides an interface for operations involving the host system such as running command files and the downloading of object data files from memory. Several important features have been recently implemented that expand the use of DISPLAY as a research tool. The program now has the capability to identify and separate the burst-friction and glottal-source sections of a sensory path into individually defined and manipulable objects. In addition, the user may now "track" along a sensory or perceptual path with a cursor and obtain the x, y, z and x′, y′, z′ coordinates of any point along the path as well as an average value for points in a user-determined subsection. This feature is invaluable in the choice of a target point for a particular section and the subsequent construction of a target zone from collections of such points. Hard copy plots of displayed data may now be generated with a six-pen plotter or with an Apple LaserWriter. Such plots may be used for journal quality reproductions of auditory-perceptual data.

SLICER is a program used in the construction of wireframe target zones that surround point data in the three-dimensional auditory-perceptual space. It displays successive slices of target data allowing the user to draw delimiting outlines around the two-dimensional slice of a target zone. This is a computerized version of the method of serial sections that has been usefully applied in microanatomical studies for many years. The vector lists which comprise the slice traces are converted to raster scans by the program CNTSYB. The rasterized data are then contoured into a three-dimensional wireframe model that represents the target zone by the commercial package SYBYL. The PS300 code which represents the target zone is then compressed by the program VCOMPRESS which reduces the amount of storage required for the target zone by as much as 80 percent.

Additionally, since a great part of the work preliminary to plotting on the Evans and Sutherland is done using two-dimensional graphics terminals, we have developed a set of software packages, which allow us to digitize and edit waveforms as well as produce plots of all the variables involved in the auditory-perceptual theory on such terminals. First, in order to simplify and standardize the writing of software which utilizes graphics, we have developed a set of 2-dimensional graphics subroutines and compiled these into a library which we call PLOT10. This library provides a functionally complete graphics

interface to any device that can emulate the Tektronix 4010 series of terminals. This library has enabled us to develop many applications that can display graphics. It has allowed researchers whose only familiarity with computers is FORTRAN to develop graphics software without involving them in the details of sending escape sequences and cryptic address coordinates. To handle the various peculiarities of different PLOT10 emulations at run-time (as opposed to compile or link time), this graphics package utilizes a system-wide text file that describes the individual characteristics of the particular terminal type being used. In this file we store items such as terminal resolution and escape sequences for entering and exiting graphics mode. This frees the programmer from dealing with the intricacies of each particular terminal, providing some degree of device independence. This package works on all of the terminals that we have access to including DEC VT240's, MicroTerm Ergo-301's, Graphon GQ-140's, and HP2623's. Hardcopy can either be obtained by screen dumps from any of our HP2623s or we can direct the graphics package to use our LN03 laser printer for publication-quality output. These routines were meant to be called from FORTRAN, but if the proper calling conventions are maintained, they may be called from any other language.

Two other important graphics routines have been developed. These are FMPTL and VAK. FMTPL allows the user to plot the values of the sensory variables SR, SF1L, SF1H, SF2, SF3, BF2, and BF3 as a function of time or as a function of distance traveled on the corresponding perceptual path in the APS. The user may choose either a logarithmic or linear frequency scale. BF2 and BF3 are clearly distinguished from SF2 and SF3 by the use of x's rather than dots. Options to plot F0 are available and options to plot and F0 modulations are planned. FMTPL also allows the user to simultaneously plot the perceptual variables PR, PF1L, PF1H, PF2, and PF3 against time or distance. Once again one may choose a linear or log frequency scale. These programs allow the user to directly view these formant tracks and compare them to what is seen in the spectrogram, what is heard, and what is observed in the APS. A variety of cursor options are planned. The graphics package VAK is oriented to the auditory-perceptual space and the search for segmentation rules. The user may plot APS coordinates (x, y, z) or slab coordinates (x', y', z'). Either sensory or perceptual values may be selected and these may be plotted against time or distance traveled. Cursor options allow one to determine the exact values of the plotted functions at any point along the curve. Additionally one may plot distance in the APS against elapsed time or the magnitude of velocity and acceleration of the perceptual pointer in APS against time or distance, with magnification of the variables and cursor measures as options. Similarly, an index of path curvature can be plotted against time and distance. Another set of VAK options includes plotting the signed velocity of the perceptual pointer in each of the dimensions x, y, z, x', y', or z' against either time or distance. These routines now allow us to quickly evaluate a variety of variables implicated as contributing to segmentation. In addition, a third routine, MULTPA, plots sensory and perceptual paths on a two-dimensional screen, along with as many target zones as the user specifies. Options include front view of the vowel space or sideview, line vs discrete symbols for each data point, and dumping to the laser printer for publication-quality output. Future work will add intensity and pitch information to the battery of plots.

We also developed our own digitization and waveform-editing package named SINS. SINS is an interactive graphical editor designed to work with a

DigiSound-16 system connected to a MicroVAX II using a SAP interface along with a DRV11-WA. SINS is an acronym for Speech IN the auditory perceptual Space. It is used for controlling analog-to-digital (A/D) and digital-to-analog (D/A) operations, as well as performing simple editing and windowing operations upon sampled waveforms. Currently we are using SINS to digitize our audio tapes recorded on our JVC VCR in our anechoic chamber. SINS is capable of reading and writing many different file formats including files which are compatible with ILS, the commercially available signal processing package that we are currently using. We have tested SINS with as many as 16 users logged onto the system at once, indicating that it is feasible to do many real-time operations on a multi-user computer running the VMS operating system. The software was written in a modular fashion so that SINS never accesses the Digisound directly. All I/O for the DigiSound-16 system is performed through the Digisound-16 library which we have developed. There are only three SINS commands that call routines from the DigiSound-16 library: play, record, and setting the sampling rate. To enable this package to work with a a different D/A-A/D system would simply require a rewrite of these three routines. SINS will provide a graphical interface, if the user is using a Tektronix 4010 compatible terminal. All graphics operations are performed using the PLOT10 library, allowing this software to be used on any type of terminal supported by the PLOT10 library. To enable the graphics to work with a different type of terminal, would simply require modifying the PLOT10 package. This should allow this software to be ported in the future to other terminal types. All other screen I/O (user input and prompting) is performed using the standard DEC Screen Management routines (SMG$ Run Time Library).

Finally, we developed a program to assist the user in editing a file which contains a list of the formants (an FMT file). The FMT file is obtained by running the program GETFIF on an analysis file which has undergone an API and a SGM (Analysis commands of the Interactive Laboratory System package). The program, named INTER, is used mainly to correct inaccuracies in our formant tracking. Of the many options, geometric interpolation, and linear interpolation are used quite frequently. Additionally, INTER can calculate the values for F1L and F1H for segments containing a voice bar. Also, columns of formant values can be copied into other columns, since a common problem with our current formant tracking is the mislabeling of formants (i.e. when F2 and F1 merge, the values placed in F2 really are values for F3). This is necessary since we do not have access to an editor that has select and paste operations on columns.

We have also implemented the Klatt synthesis program on our MicroVAX II. This program has been modified in several ways by adding subroutines that enhance the front end. We now have options for different input glottal waveforms and output directly to an ILS file. We can now also use a digitizer pad with the front view of the vowel slab to enter x', y' coordinates with a pre-set z'. These values are then automatically converted to formant values and bandwidths which are used as parameters for synthesis. A separate program has been written which allows batch synthesis overnight of great numbers of stimuli without requiring the presence of the experimenter. This capability will now be used to precisely define the borders of the target zones.

These packages, all of them developed in the last two years, provide an excellent environment for carrying out our research. We now plan to enhance the

software, as we keep developing the auditory-perceptual theory, so that the end result will be a hands-off processing of the acoustical signal of speech. All of these programs are necessary to enable us to conduct our basic research on human speech perception.

Software for the VAXStation 8000 and the Evans and Sutherland PS300. A DEC VAXStation 8000 color graphics workstation, housed in the Research Department at CID, was recently obtained with a grant from the Southwestern Bell Foundation. We are in the process of converting all of our graphics programs currently implemented on the PS330 to this machine. With its virtual memory capability and high-speed bus, the VAXStation 8000 will permit storage of many more graphics objects in memory as well as greatly reduce the time it takes to download these objects from data files. In the meantime, we are still working mainly with the PS330, and in the period September 1987 through September 1988 the following major efforts were completed in the area of graphics sofware development:

a) The capability of separating displayed sensory paths into glottal-source and burst-friction segments was added.

b) A cursor function, which allows the user to step through a sensory or perceptual path by means of a graphics cursor, and to calculate and display various metrics associated with points along the path, was added.

c) A facility for obtaining a black and white hard-copy of a displayed screen, via a postscript laser printer, was added.

d) A separate program, SLICER, which enables the user to create interactively a three-dimensional target zone, was significantly enhanced and then incorporated into DISPLAY, the primary software tool for the analysis of speech on the Evans and Sutherland computer.

e) An animation package was developed for the PS330 (Evans and Sutherland Corporation) computer that allows the user to submit a plain language script and translates this script into graphics and control commands for the filming of a particular sequence.

D. Projects completed and published or submitted for publication. (Joint AFOSR & NIDCD Funding)

1. Miller, J.D. (1987). "Auditory-perceptual processing of speech waveforms," in Auditory Processing of Complex Sounds, edited by W. A. Yost and C. S. Watson (Erlbaum, Hillsdale, NJ) 257-266.

   An account of the processes whereby the acoustic waveform of speech is converted by the human listener to a representation that is isomorphic with a sequence of allophones is presented. This account is based on the author's auditory-perceptual theory of phonetic recognition (Miller, 1984abc). Three stages of processing are identified. In Stage I, the acoustic waveform is converted into sensory variables that represent the short-term spectral patterns associated with the waveform as well as their loudnesses and goodnesses. A key notion is that these patterns are represented as points in an phonetically relevant auditory-perceptual space, which is usually conceived as having three dimensions. In Stage II, these variables are integrated by a sensory-perceptual transformation into a single, unitary response. This perceptual response (perceptual pointer) can also be represented at any moment as a point in the auditory-

perceptual space, and over time a sequence of points or a perceptual path is generated. Stage III is the perceptual-linguistic transformation. Here the dynamics of the perceptual pointer in relation to perceptual target zones within the auditory-perceptual space cause those target zones to issue category codes or neural symbols that are isomorphic with the allophones of the language. In a fourth stage, not dealt with here, the sequence of category codes so generated is converted to units isomorphic with the language's lexicon. In this paper, the auditory-perceptual space is described and some of the characteristics of the preliminary estimates of the target zones are described. Also, the hypothesized sensory-perceptual transformation is described as is a possible segmentation maneuver.

2. Fourakis, M. and Monahan, C.B. (1988). "Effects of metrical foot structure on syllable timing," Language and Speech, 31, 283-306.

   The durations of syllabic intervals in sentences with different rhythmic structure were examined. Rhythmic structure was defined as the organization of stressed and unstressed syllables into metrical feet - in this case, iambs and anapests. Each sentence contained three metrical feet, and each foot could be either an iamb or an anapest; hence, there were eight sentences of different rhythm types. Six native speakers of American English each read 24 versions of each of the 8 rhythm types (192 sentences). The utterances were recorded and the durations of each syllable of each foot were measured from sound spectrograms. Statistical analysis indicated that the durations of some syllables were affected by the structure of the foot that contained them. There was also, in most cases, a significant shortening of the stressed syllable of a foot when the following foot was an anapest rather than an iamb. The results were interpreted as indicating that the process by which speech is produced is not strictly concatenative but involves planning which extends to at least two metrical feet at a time.

3. Jongman, A., Fourakis, M. and Sereno, J. (1989). "The acoustic vowel space of Modern Greek and German," Language and Speech, 32, 221-248.

   The spectral characteristics of vowels in Modern Greek and German were examined. Four speakers of Modern Greek and three speakers of German produced four repetitions of words containing each vowel of their native language. Measurements of the fundamental frequency and the first three formants were made for each vowel token. These measurements were then transformed into log frequency ratios and plotted as points in the three-dimensional auditory-perceptual space proposed by Miller (1989). Each vowel token was thus represented by one point, and the points corresponding to each vowel category were enclosed in three-dimensional target zones. For the present corpus, these zones differentiate the five vowels of Modern Greek with 100% accuracy, and the fourteen vowels of German with 94% accuracy. Implications for the distribution of common vowels across languages as a function of vowel density are discussed.

4. Miller, J.D. (1989). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. 85, 2114-2134.

The major issues in relating acoustic waveforms of spoken vowels to perceived vowel categories are presented and discussed in terms of the author's auditory-perceptual theory of phonetic recognition. A brief historical review of formant-ratio theory is presented, as well as an analysis of frequency scales that have been proposed for description of the vowel. It is illustrated that the monophthongal vowel sounds of American English can be represented as clustered in perceptual target zones within a three-dimensional auditory-perceptual space (APS), and it is shown that preliminary versions of these target zones segregate a corpus of vowels of American English with 93% accuracy. Furthermore, it is shown that the nonretroflex vowels of American English fall within a narrow slab within the APS, with spread vowels near the front of this slab and rounded vowels near the back. Retroflex vowels fall in a distinct region behind the vowel slab. Descriptions of the vowels within the APS are shown to be correlated with their descriptions in terms of dimensions of articulation and timbre. Additionally, issues related to talker normaliation, coarticulation effects, segmentation, pitch, transposition, and diphthongization are discussed.

5. Miller, J.D. and Fourakis, M. (1989). "Speech perception," International Encyclopedia of Communication.

6. Fourakis, M.S. "Stress, tempo, and vowel reduction in American English." To be submitted to J. Acoust. Soc. Am.

Two processes that affect the acoustic characteristics of vowels are discussed, phonological and phonetic vowel reduction. Phonological vowel reduction applies to unstressed vowels. Phonetic vowel reduction is supposed to apply to all vowels and be caused by fast speech rates, context, as well as lack of stress. In this experiment, the effects of changes in stress and in rate of speech (tempo) on the acoustic characteristics of American English monophthongal, nonretroflex vowels were examined. Four male and four female native speakers produced these vowels in two contexts, [h_d] and [b_d], in a carrier sentence, under four conditions of tempo-stress (slow-stressed, slow-unstressed, fast-stressed, and fast-unstressed). A total of 2304 vowel utterances were collected (8 speakers x 9 vowels x 4 repetitions x 2 contexts x 4 tempo-stress conditions). Measurements of duration and F0 showed that the subjects did in fact vary tempo and stress as instructed. The effect of a change in stress on vowel duration was found to be slightly larger than that of a change in tempo. The putative vowel portion of each utterance was analyzed, formant tracks were obtained, and these were plotted in an auditory-perceptual space [Miller, J. Acoust. Soc. Am. 85, 2114-2134 (1989)]. These plots served to determine the part of the utterance that could, in most cases, be considered its steady state, and could be represented by a point in the space, the coordinates of which were given by the average, over time, of the coordinates of the steady-state portion. The distance of these data points from the point representing the acoustic characteristics of a vowel produced by a neutral vocal tract was used to determine the magnitude of phonetic vowel reduction caused by faster tempo and less stress, relative to the slow-stressed condition. The results indicate that these distances do not have a major dependence on tempo and stress. In addition, several vowel classifications schemes were tested

using linear discriminant analysis, and the one proposed by Miller (1989) performed better than combinations of F0, F1, F2, and F3.

7. Hawks, J.W. (1990). "Perceptual aspects of a three-dimensional vowel space," Ph.D. dissertation, Washington University, St. Louis, MO.

One of the methods that has been utilized for vowel classification in natural speech posits a three-dimensional space, defined by certain acoustic characteristics of the vowels related to the first three significant prominences, or formants, in their short-term spectra (F1, F2, and F3) and voice pitch (F0). Within this space, productions of like vowels are mapped onto subspaces, called target zones. Here, the characteristics of this space and its subspaces were studied by means of subject responses to synthetic vowel-like tokens representing unique points in this space. The first experiment utilized the identifications of eight listeners to construct a vowel map of this space. Vowel categories for American English can be represented as abutting and non-overlapping target zones which correctly classify over 99% of the plurality-based identifications. The first two formants (F1 and F2) appeared to be the primary determinants in the identification of non-retroflex vowels. The third formant (F3) determined the perception of retroflex (r-colored) vowels, and contributed to the phonetic saliency of other vowel categories. Furthermore, phonetic saliency varied in an orderly way with location within a vowel target zone.

A second experiment investigated the discrimination of complex sounds represented as points in the three-dimensional vowel space by estimating difference limens (DLs), expressed as distance in the space, for synthetic, vowel-like tokens. Four subjects were used to estimate DLs along 102 straight-line continua. Continua emanated in six directions from 17 locations in the space. Movement along continua resulted in distinct, multi-formant patterns of frequency change which varied with the direction and axis of movement. The average DL for distance across all continua was estimated to be .01 log units, but DLs were significantly different for the three axes of the space. Considerable variation found in DLs associated with individual reference points may be related to differences in reference formant patterns. The DL results for multiple-formant-change continua were found to be significantly smaller than similar DL estimates for single-formant-change continua. Many of these differences could be accounted for by an additive model whereby each changing formant contributes independent information to perception.

E. Projects completed and reported at scientific meetings. (Joint AFOSR & NIDCD Funding)

1. Chang, H.M. (1987). "Automatic detection for diphthongs," J. Acoust. Soc. Am., 82(S1), S37(A).

An account of the processes whereby the acoustic waveform of speech is converted by the human listener to a representation that is isomorphic with a sequence of allophones is presented. This account is based on the author's auditory-perceptual theory of phonetic recognition (Miller,

1984abc). Three stages of processing are identified. In Stage I, the acoustic waveform is converted into sensory variables that represent the short-term spectral patterns associated with the waveform as well as their loudnesses and goodnesses. A key notion is that these patterns are represented as points in an phonetically relevant auditory-perceptual space, which is usually conceived as having three dimensions. In Stage II, these variables are integrated by a sensory-perceptual transformation into a single, unitary response. This perceptual response (perceptual pointer) can also be represented at any moment as a point in the auditory-perceptual space, and over time a sequence of points or a perceptual path is generated. Stage III is the perceptual-linguistic transformation. Here the dynamics of the perceptual pointer in relation to perceptual target zones within the auditory-perceptual space cause those target zones to issue category codes or neural symbols that are isomorphic with the allophones of the language. In a fourth stage, not dealt with here, the sequence of category codes so generated is converted to units isomorphic with the language's lexicon. In this paper, the auditory-perceptual space is described and some of the characteristics of the preliminary estimates of the target zones are described. Also, the hypothesized sensory-perceptual transformation is described as is a possible segmentation maneuver.

2. Fourakis, M.S. and Miller, J.D. (1987). "Measurements of vowels in isolation and in sonorant context," J. Acoust. Soc. Am., 81(S1), S17(A).

   In Miller's auditory-perceptual theory vowels are hypothesized to occupy nonoverlapping zones in a three-dimensional space defined by the relative positions of the first three spectral prominences of the short-term spectrum and by a reference related to the speaker's characteristics. In order to validate target zones drawn on the basis of the data available in the literature, we have recorded two male and two female speakers of American English producing the nine pure vowels of English in isolation and in sonorant context. The recordings were digitized at 20,000 Hz with 14-bit precision and analyzed using ILS. The formant and fundamental frequency information was extracted from the ILS analysis files. The data are now in the process of being smoothed in the log-frequency domain and are being transformed into paths in the three-dimensional auditory perceptual space. By examining not only the pure vowels but also vowels in sonorant context, the aim is to establish the coarticulatory effects on relative spectral prominence positions of preceding and following [1] and [2]. [Work supported by NINCDS and AFOSR.]

3. Hawks, J.W. and Miller, J.D. (1987). "Listener-talker interaction: Is there an 'autophonetic' effect?" J. Acoust. Soc. Am., 81(S1), S4.

   Based on recent physiological evidence for neural selectivity to self-produced song in the auditory system of the White-Crowned Sparrow, a hypothesis stating that the human perceptual mechanism for speech may be most acute for self-produced speech sounds was tested. A test of the ability to identify monosyllabic words was presented in a noise background. The same subjects served as both talkers and listeners. A small, but statistically nonsignificant, advantage is found when listener and talker are the same person. Additionally, while variation between

talkers was nearly twice the variation between listeners, the performance of a particular talker-listener pair could be predicted by a simple equation. This equation gives equal weight to the talker's and listener's average performance, and thus the talkers, by virtue of their greater variance, tend to have a greater influence. [Work supported by NIH.]

4. Miller, J.D. (1987). "Classification of vowel productions by means of perceptual target zones: A response to Ladefoged and Studdert-Kennedy." J. Acoust. Soc. Am., 82(S1), S82(A).

   In discussion of my paper [Miller, J. Acoust. Soc. Am. Suppl. 81, S16 (1987)] at the 113th Meeting, Dr. Ladefoged and Dr. Studdert-Kennedy requested specific numbers regarding percent correct classifications of vowel productions achievable through use of the auditory-perceptual theory and its perceptual target zones (PTZs). At the time of the 113th Meeting, the estimated shapes and locations of the PTZs were very preliminary and represented only a second iteration (I2). The percents correct for the I2 zones were moderate (50%-95%). Based on reconsideration of older data and on numerous new measurements made in our laboratory, a new set of PTZs is being developed for the vowels. These represent the third iteration or I3 zones. The I3 zones will be presented and percentages of correct classifications for several data sets will be reported. Issues involved in developing decisive tests of the perceptual target zones will be discussed. [Supported by NINCDS and AFOSR.]

5. Gottfried, M. (1989) "Some acoustical properties of diphthongs," J. Acoust. Soc. Am., 86, S123 (ASA meeting, Fall 1989).

   One approach to describing diphthongs is to consider them as consisting of a steady-state nucleus and a glide toward another steady-state segment. This study evaluates this approach when sources of acoustic variation (change of speaker, stress, and rate) are introduced. In a preliminary study, four speakers of Midwestern American English were recorded producing diphthongs in isolation and in the context /K_T/. On the basis of the first three formants, diphthongs can indeed be characterized in terms of two kinds of spectrally defined states, steady-state and glide. There is, however, variation in the number and order of these states for particular diphthongs. Plots of these spectral patterns in an auditory-perceptual space [J.D. Miller, J. Acoust. Soc. Am. 85, 2114-2134 (1989)] suggested that diphthongs can be differentiated by a combination of distinctive nucleus zones and angular movements within the space. Productions by two speakers of a carrier sentence containing diphthongs under varied stress and rate conditions are being analyzed. Results will be used to evaluate the usefulness of the descriptive scheme noted above and the ability of the auditory-perceptual space to capture differences between diphthongs. Results of this evaluation, in addition to durational measurements of individual diphthongs and their component segments, will be reported. [Work supported by NINCDS.]

6. Hawks, J.W. (1989) "Perception of synthetic vowels: A comparison of several classification schemes," J. Acoust. Soc. Am., 86(S1), S78 (ASA meeting, Fall 1989).

Synthetic vowel tokens (1725) were randomly presented twice to eight native speakers of Midwestern American English for classification as one of 12 vowel categories, /IY,IH,EH,EY,AE,AA,AH,AO,OW,UH,UW,ER/. Subjects rated the certainty of their responses on a scale from one (very unsure) to five (very sure). The vowels were synthesized in null context and utilized a male F0 contour. The frequency values of F1, F2, and F3 were selected such that the entire area where vowels may be represented in the auditory-perceptual space [J.D. Miller, J. Acoust. Soc. Am. 85, 2114-2134 (1989)] was equidistantly sampled. The results of this experiment support the view that F1 and F2 are the primary determinants in the perception of nonretroflex vowels. In addition, monophthongal versions of the diphthongs /EY/ and /OW/ may be distinctly calssified. The results will be presented graphically as target zones constructed on the basis of the plurality identifications for each token. These target zones are abutting and nonoverlapping, and correctly classify 99.9% of the plurality judgments (less 48 tokens where ties occurred) and 75% of the 27,600 total judgments. Other vowel classification schemes will also be compared for their accuracy in classifying the results of this experiment, as well as results from other studies. [Work supported by NINCD.]

7. Miller, J.D. and Hawks, J.W. (1989). "Target zones for synthetic vowels," J. Acoust. Soc. Am., 85(S1), S51 (ASA meeting Spring 1989).

Using algorithms developed by the second author (JWH), he synthesized 7703 vowel tokens such that the values of F0, F1, F2, and F3 uniformly sampled locations in a plane in the middle of the vowel slab of the auditory-perceptual space [J.D. Miller, "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. (in press)]. Each token was 400 ms and had a rise-fall pitch contour. The tokens were arranged in random order and presented in groups of 100 to one listener (JDM), who identified each token as [IY,IH,EH,EY,AE,AA,AH,AO,OW,UH, or UW] and rated the clarity of each token on a scale from 1 (poor) to 5 (excellent). The sizes and shapes of the target zones derived from these experiments will be compared to those based on measurements of natural speech. The data are, in general, consistent with the concept of vowel target zones as being large with irregular and abutting boundaries. [Work supported by NINCDS.]

IV. OTHER RESEARCH BY PARTICIPANTS

A. Publications (only references without abstracts).

1. Boettcher, F. and Miller, J.D. (1986). "The formation of learning sets by the chinchilla in an auditory discrimination task," J. Auditory Res., 26, 99-113.

2. Fourakis, M. (1986). "An acoustic study of the effects of tempo and stress on segmental intervals in modern Greek," Phonetica, 43, pp. 172-188.

3. Fourakis, M. (1986). "A timing model for word initial CV-syllables in Modern Greek," J. Acoust. Soc. Am., 79, pp. 1982-1986.

4. Fourakis, M. and Port, R. (1986). "Stop epenthesis in English," J. Phonetics, 14, 197-221.

5. Skinner, M.W., Miller, J.D., DeFilippo, C.L., Dawson, J.K., and Popelka, G.R. (1986). "Word identification by listeners with sensorineural hearing loss using four amplification systems," in: M.J. Collins, T.J. Gattke, and L.A. Harker (Eds.) Sensorineural Hearing Loss, Iowa City, Iowa: University of Iowa, pp. 305-325.

6. Bowe, C.A., Miller, J.D., and Green, L. (1987). "Qualities and locations of stimuli and responses affecting discrimination learning of chinchillas (Chinchilla laniger) and pigeons (Columba livia)," J. Comp Psych., 101 No.2, pp. 132-138.

7. Fourakis, M. and Iverson, G.K. (1987). "More on second-language learning," Language Learning, 37:297-302.

8. Osberger, M.J., Johnson, D.L., and Miller, J.D. (1987). "Use of connected discourse tracking to train functional speech skills," Ear & Hearing, 8, pp. 31-36.

9. Sereno, J.A., Baum, S.R., Marean, G.C., and Lieberman, P. (1987). "Acoustic analyses and perceptual data on anticipatory labial coarticulation in adults and children," J. Acoust. Soc. Am., 81, 512-519.

10. Sereno, A.J. and Lieberman, P. (1987). "Developmental aspects of lingual coarticulation," J. of Phonetics, 15, 247-257.

11. Weisenberger, J.M., Heidbreder, A.F. and Miller, J.D. (1987). "Development and preliminary evaluation of an earmold sound-to-tactile aid for the hearing-impaired," J. Rehab. Research and Development, 24, pp. 51-66.

12. Weisenberger, J.M. and Miller, J.D. (1987). "The role of tactile aids in providing information about acoustic stimuli," J. Acoust. Soc. Am., 82 (3), pp. 906-916.

13. Sereno, J.A. (in press). "Phonosyntactics," In: L. Hinton, J. Nichols, and J. Ohala (Eds.), Proceedings from a Conference on Sound Symbolism, Cambridge: Cambridge University Press.

V. PARTICIPATING PROFESSIONALS

James D. Miller

University of Wisconsin, Madison, WI	B.S.	1951	Psychology
Indiana University, Bloomington, IN	M.A.	1953	Psychology
Indiana University, Bloomington, IN	Ph.D.	1957	Psychology

Dissertation title: "On the relationship between temporary hearing loss and masking."

Marios S. Fourakis

Wabash College, Crawfordsville, IN	B.A.	1973	Classical Greek Lit.
Indiana University, Bloomington, IN	M.A.	1979	Classical Studies
Indiana University, Bloomington, IN	M.A.	1980	Linguistics
Indiana University, Bloomington, IN	Ph.D.	1983	Linguistics

Dissertation title: "An acoustic study of temporal programming in speech production."

Michael Gottfried

University of Pennsylvania, PA	B.A.	1976	Linguistics
University of Pennsylvania, PA	M.A.	1976	Linguistics
University of Pennsylvania, PA	Ph.D.	1986	Linguistics

Dissertation title: "Cross-reference in a scientific sublanguage."

Allard Jongman

University of Amsterdam, The Netherlands	B.A.	1980	Slavics
University of Amsterdam, The Netherlands	M.A.	1982	Linguistics
Brown University, Providence, RI	Ph.D.	1986	Linguistics

Dissertation title: "Naturalness in phonetics: A study of context-dependency."

Frank E. Kramer

University of Houston, Houston, TX	B.S.	1972	Geology
Washington University, St. Louis, MO	M.S.	1987	Geochemistry

Thesis title: "The distribution of Krypton in an anorthite-diopside-water mixture at five kilobars pressure."

Steven J. Sadoff

Washington University, St. Louis, MO	B.S.	1985	Computer Science
Washington University, St. Louis, MO	M.S.	1987	Electrical Eng.
Washington University, St. Louis, MO	Sc.D.	1990	Computer Science

Dissertation title: "Classifying speech into silence, glottal source, burst friction, or mixed categories."

Joan A. Sereno

Northern Illinois University, De Kalb, IL	B.S.	1982	Psychology
Northern Illinois University, De Kalb, IL	B.A.	1982	Philosophy
Brown University, Providence, RI	M.A.	1986	Linguistics

Brown University, Providence, RI Ph.D. 1987 Linguistics

Dissertation title: "Graphemic, associative, and syntactic priming effects at a brief stimulus onset asynchrony in naming and lexical decision."

Lynn W. Shields

University of Tennessee, Knoxville, TN M.A. 1976 Speech Pathology

Mark R. Veksler

Washington University, St. Louis, MO B.S. 1986 Systems Science &Math

LaDeana F. Weigelt

Northwest Nazarene College, Nampa, ID B.S. 1986 Engineering Physics
Northwestern University, Evanston, IL M.S. 1988 Biomedical Eng.

Bee-Kong Yew

Southern Illinois University, Carbondale, IL B.S. 1986 Bus.Adm./Mgmt.
Southern Illinois University, Carbondale, IL M.B.A. 1987 Bus.Adm./Mgmt.

Appendix

Final Technical Report
for AFOSR Grant G-AFOSR-86-0335
"Auditory-Acoustic Basis of Consonant Perception"

Contents

I. Doctoral Dissertations

A. Chang, H.M. (1987). "SWIS: See what I say, a speaker-independent word recognition system by phoneme-oriented mapping on a phonetically encoded speech map," D.Sc. Dissertation, Washington University, St. Louis, MO, pp. 254.

B. Sadoff, S.J. (1990). "Classifying speech into silence, glottal source, burst friction, or mixed categories," D.Sc. Dissertation, Washington University, St. Louis, MO, pp. 114.

C. Hawks, J.W. (1990). "Perceptual aspects of three-dimensional vowel space," Ph.D. Dissertation, Washington University & Central Institute for the Deaf, St. Louis, MO, pp. 221.

II. Reprints

D. Miller, J.D. (1989). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. 85, 2114-2134.

E. Jongman, A., Fourakis, M., and Sereno, J. (1989). "The acoustic vowel space of modern Greek and German," Lang. and Speech 32, 221-248.

F. Weigelt, L.F., Sadoff, S.J., and Miller, J.D. (1990). "Plosive/fricative distinction: the voiceless case," J. Acoust. Soc. Am. 87, 2729-2737.

III. Manuscripts

G. Weigelt, L.F., Sadoff, S.J., Miller, J.D. (Submitted to Speech Communication). "Plosive/fricative distinction: the voiced case," pp. 39.

H. Fourakis, M. (Submitted to J. Acoust. Soc. Am.). "Stress, tempo, and vowel reduction in American English," pp. 51.

I. Gottfried, M. and Miller, J.D. (Submitted to Journal of Phonetics). "An approach to the classification of American English diphthongs," pp. 69.

To Dr. James Miller

ENGR 854, 853

Hisao Ming Chang 5/14/87

SEVER INSTITUTE OF TECHNOLOGY

Doctor of Science Degree

AFOSR-TR- 91 0130

DISSERTATION ACCEPTANCE
(To be submitted by the graduation approval deadline)

DATE: December 15, 1986

STUDENT'S NAME: Hisao Ming Chang

E.R.S. CODE: _____

This student's dissertation, entitled "SWIS: See What I Say -- A Speaker-Independent Word Recognition System by Phoneme-Oriented Mapping on a Phonetically-Encoded Auditory-Perceptual Speech Map"

has been examined by the undersigned committee of three faculty members and has received full approval for acceptance in partial fulfillment of the requirements for the degree Doctor of Science.

Signatures: James D. Miller Co-Chairman

Jerome R. Cox Co-Chairman

Robert E. Nickell

Distribution:

5 - Dissertation copies
1 - Candidate
1 - Department
1 - Dean's Office
1 - Registrar

AFOSR Grant G-AFOSR-86-0335
Final Technical Report
Appendix

A

WASHINGTON UNIVERSITY
SEVER INSTITUTE OF TECHNOLOGY

ABSTRACT

SWIS: SEE WHAT I SAY

A SPEAKER-INDEPENDENT WORD RECOGNITION SYSTEM BY

PHONEME-ORIENTED MAPPING ON A PHONETICALLY-ENCODED

AUDITORY-PERCEPTUAL SPEECH MAP

By Hisao (Harry) M. Chang

ADVISOR: Professor J. R. Cox, Jr.

May, 1987

Saint Louis, Missouri

This paper reports the development of a prototype speaker-independent word recognition system based on a new model of phoneme perception proposed in the auditory-perceptual theory of phonetic recognition (see Appendix A). The system is able to map directly acoustical parameters to phonetic events by means of a phonetically-encoded auditory-perceptual map designed for a selected set of phonemes derived from all the stressed syllables in an 11-word digit vocabulary (0 through 10). When the system is tested using a 231 token database recorded from 21 new talkers (11 males and 10 females) for the digit vocabulary, a 61.5% score for phoneme recognition is obtained while the phoneme insertion rate is 4.25%. A 53% score is obtained for word recognition by using only the phonetic information derived from the vocalic segments of a stressed syllable in a word utterance.

After reviewing the sources of success and failure it is argued that a phoneme-oriented system based on the broad framework established in the theory and implementated in this project, could be generalized to provide a solution to speaker-independent isolated word recognition for a medium-sized vocabulary of 50 to 200 words.

TABLE OF CONTENTS

NO.		Page
1.	Introduction	1
	1.1 History and Motivation	1
	1.2 Background	5
	1.2.1 Isolated-Word Recognition	6
	1.2.2 Continuous-Speech Recognition	7
	1.2.3 Speaker Independence	10
	1.3 Problem Areas	12
	1.3.1 Recognition Unit	12
	1.3.1.1 Word/Phrase Template	12
	1.3.1.2 Syllable-Template	13
	1.3.1.3 Demisyllables	15
	1.3.1.4 Phoneme	16
	1.3.2 Segmentation	17
	1.3.3 Feature Extraction	19
	1.3.4 Cues to Phonetic Identities	22
	1.4 Literature Review	23
	1.5 The Research Plans and Goals	24
	1.5.1 The Project Scope	25
	1.5.2 Acoustic Data Acquisition and the Database Used	28
	1.5.3 Limitation and Research Source	30
2.	The System Concept and SWIS Overview	34
	2.1 Introduction	34
	2.2 System Working Assumptions and SWIS Overview	36
	2.2.1 Comparative Discussion	36
	2.2.2 System Working Assumption	39
	2.2.3 SWIS Overview	40

TABLE OF CONTENTS
(continued)

No.		Page
3.	Signal Processing and Feature Representation	43
	3.1 Acoustic Preprocessor and Linear-Prediction Analysis (LPA)	43
	3.2 Feature Representation	43
	3.2.1 Fundamental Frequency (FO)	45
	3.2.2 Sensory Reference (SR)	46
	3.2.3 First Spectral Peak	46
	3.2.4 Maximum Energy Peak	47
	3.2.5 Smoothed Sensory Formants	49
	3.2.6 Perceptual Path	49
	3.2.7 Velocity and Acceleration	50
	3.2.8 Segmentation Index	51
	3.2.9 Angle and Duration	52
4.	SWIS Front-End	57
	4.1 Introduction	57
	4.2 Classification	59
	4.2.1 Glottal-Source Detection: **ALGORITHM, G**	61
	4.2.2 Sonorant-Energy Dip: **ALGORITHM, SED**	67
	4.2.3 Nasal Segmentation: **ALGORITHM, N**	72
	4.2.4 Discussions	79
	4.3 Sensory Path Generation	81
	4.3.1 Introduction	81
	4.3.2 Review of Formant Extraction Methods	83
	4.3.3 Spectrum Representation	85
	4.3.4 Identification of Sensory Formants	88
	4.3.5 Performance of the SPG Algorithm	95
5.	Phonetic Decoder	99
	5.1 Introduction	99
	5.2 Sensory-Perceptual Transformation	100

TABLE OF CONTENTS
(continued)

No.		Page
	5.3 Phonetic-Encoded Auditory-Perceptual Map	104
	5.4 Critical Segment Detection	110
	5.4.1 Anchor Frame	113
	5.4.2 Pre-Critical Segment	113
	5.4.3 Post-Critical Segment	114
	5.4.4 Tail-Critical Segment	116
	5.4.5 Phonetic Code Generation	118
	5.5 Diphthong Recognition	121
	5.6 Phonetic Dictionary	127
	5.7 Word Generation	132
6.	Experimental Evaluation of SWIS Performance	136
	6.1 Classification Accuracy	136
	6.2 Accuracy of Phoneme Recognition	141
	6.3 Word Recognition	146
	6.4 Discussion	149
7.	Implementation	152
8.	Conclusion	157
9.	Acknowledgments	159
10.	Appendices	160
	Appendix 10.1 The Auditory-Perceptual Theory of Phonetic Recognition	161
	Appendix 10.2 Classification PLots for Every Vocabulary Word	188
	Appendix 10.3 Listing of Sensory Formant Generation Errors Output from an Error Analysis Program	211

TABLE OF CONTENTS
(continued)

No.		Page
	Appendix 10.4 Sample Plots of Smoothed Sensory Formants ...	219
11.	Bibliography	242
12.	Vita	254

LIST OF TABLES

No.		Page
1.1	ARPAbet Symbols for Representing within a Computer the Phoneme-like Units of English	3
1.2	Digit Vocabulary	26
1.3	Phonemes Studied in SWIS	26
1.4	Data Corpuses Used	29
3.1	Feature Listing	44
4.1	Lexical Presentation of DIGIT in Broad Phonetic Classes	62
4.2	Classification Errors Computed from the Training Database TRAIN	80
4.3	Classification Errors Computed from the Training Database TEST1	80
5.1	Phonetic Labels Used in SWIS Phonetic Dictionary	128
5.2	SWIS Standard Phonetic Dictionary	131
5.3	SWIS Advanced Phonetic Dictionary	132
6.1	Distributions of Classification Errors	138
6.2	Word Recognition Accuracy	146
7.1	Sample Phonetic Code Sequences Generated from the Test Database TEST2	153

LIST OF FIGURES

No.		Page
1.1	The Sample Plots of the Speech Waveforms Showing the Irregularity of "Pulse Chains"	20
1.2	An Illustration for the Problem "Peak-Picking" in Formant Estimation. The Utterance "SEVEN" is made by an Adult Female.	20
2.1	Illustrations of a set of Burst-Friction Spectra as in (a) and a set of Glottal-Source Burst-Friction Mixed Spectra as in (b)	37
2.2	The Major Components of SWIS Recognition System	41
3.1	Illustration of Selecting P1 and PM in Different Phonetic Contexts	48
3.2	Illustration of a Y'-glide and its Attributes. The Y'-glide Starts at ith Frame and Ends at jth Frame (Darken Segment)	54
4.1	The Block Diagram of the SWIS Front-End System	58
4.2	Illustration of Multiple Features-Based Phonetic Classification	64
4.3	Various Appearances of Sonorant-Energy Dip due to Syllabic Contexts and Talkers	68
4.4	Mean Error Rates Computed over each Formant Contour of 10 Words in DIGIT, two Utterances per Word (one Female Talker and one Male Talker)	98
5.1	Effects of Using the two Different Center Frequencies in the Sensory-Perceptual Transformation Function for the same Input Variable	102
5.2	Sketch of the Phonetically-Encoded Auditory-Perceptual Map	109
5.3	Time Variations of the first two Formants for six Diphthong Plots (After Holbrook and Fairbanks).....	122
5.4	The Perceptual Paths of some Diphthong Segments are Plotted on the P-map to Illustrate the Concepts of Diphthong Recognition.....	124

LIST OF FIGURES
(continued)

No.		Page
6.1	Error Rates Grouped by their Types are Computed from all the Utterances in the Testing Database TEST2.	140
6.2	Recognition Rates and Insertion Rates Plotted for each Phoneme Studied in SWIS are Based on the Test Results of all the Utterances in the Database TEST2.	144

SWIS: SEE WHAT I SAY

A SPEAKER-INDEPENDENT WORD RECOGNITION SYSTEM BY PHONEME-ORIENTED MAPPING ON A PHONETIC-ENCODED AUDITORY-PERCEPTUAL SPEECH MAP.

1. INTRODUCTION

1.1 HISTORY AND MOTIVATION

The research on Automatic Speech Recognition (ASR) by machine has entered a highly transitional stage in recent years. In the last decade most research efforts focused on the development of recognition systems that employ mature general-purpose pattern recognition techniques but utilize little or no speech-specific knowledge. The success of these systems can be attributed mainly to the use of semantic, syntactic and discourse constraints in well-constrained speech recognition tasks. The DARPA SUR project (from 1971 to 1976) represented some of the best work ever done in the development of these "top-down" systems (1-5).*

However, it has generally been recognized that the extendibility of these techniques to tasks involving multiple speakers, large vocabularies, continuous speech, and less-restricted language is very limited (6-8). In the late 1970s and at the beginning of this

* The numbers in parentheses in the text indicate references in the Bibliography.

decade, research efforts have gradually shifted to the development of feature-based systems because they promise substantially better performance in ASR (9-13). These systems place great emphasis on "front-end" analysis, i.e., using a "bottom-up" method to find so-called acoustic invariance. The basic assumption of these techniques is that the acoustic signal is very rich in phonetic information. The pioneering spectrogram reading experiments (14-17) strongly support this belief. Recent research projects funded by DARPA and other government agencies have clearly shown that this trend will continue to dominate the ASR research in the scientific community through the next decade (8,18-22).

One of the fundamental issues in the design of a feature-based system is the development of models in which the phonetic identities of speech sounds and their acoustic correlates can best be explored. The recent auditory-perceptual theory of Miller (23) has proposed such a model with many interesting properties. A preliminary study (24) based on this theory shows 96.4% classification accuracy for three closely-spaced vowels (/ IH EH AH /*). Another experiment based on the theory demonstrates 93% classification accuracy for 5 vowels (/IY EH AE OW UW/) (25). It is in this context of the auditory-perceptual theory that the development of a phoneme-based word recognition system is proposed to investigate both the adequacy and the feasibility of the theory. From now on we will use APT to refer to the auditory-perceptual theory of Miller.

*The ARPAbet symbol system is used in this report to provide a standard phonetic transcription, see Table 1.1 for the details.

Phoneme	Computer Representation		Example	Phoneme	Computer Representation		Example
	1-Character	2-Characters			1-Character	2-Characters	
i	i	IY	beat	p	p	P	pet
ɪ	I	IH	bit	t	t	T	ten
e	e	EY	bait	k	k	K	kit
ɛ	E	EH	bet	b	b	B	bet
æ	@	AE	bat	d	d	D	debt
ɑ	a	AA	Bob	g	g	G	get
ʌ	A	AH	but	h	h	HH	hat
ɔ	c	AO	bought	f	f	F	fat
o	o	OW	boat	θ	T	TH	thing
U	U	UH	book	s	s	S	sat
u	u	UW	boot	š or ʃ	S	SH	shut
ə	x	AX	about	v	v	V	vat
ɨ	X	IX	roses	ð	D	DH	that
ɝ	R	ER	bird	z	z	Z	zoo
aU or aw	W	AW	down	ž or ʒ	Z	ZH	azure
aI or ay	Y	AY	buy	č	C	CH	church
ɔI or ɔy	O	OY	boy	ǰ	J	JH	judge
y	y	Y	you	ʰ	H	WH	which
w	w	W	wit	syl l, l̩	L	EL	battle
r	r	R	rent	syl m, m̩	M	EM	bottom
l	l	L	let	syl n, n̩	N	EN	button
m	m	M	met	flapped t,r	F	DX	batter
n	n	N	net	glottal stop, ʔ	Q	Q	
ŋ	G	NX	sing	Silence	-	-	
				non-speech Segment	!	!	laugh, etc.

AUXILIARY SYMBOLS (1- AND 2-CHARACTER CODES ARE IDENTICAL)

Symbol	Meaning	Symbol	Meaning
-	Morpheme boundary	:3 or .	Fall-rise or non-term juncture
'	Word boundary	* **	Comment (anything except * or **)
#	Utterance boundary	' '	Apos.-surround special symbol in comment
	Tone group boundary	( )	Phoneme class information
:1 or `	Falling or decl. juncture	< >	Phonetic or allophonic escape
:2 or "	Rising or inter. juncture		

STRESS REPRESENTATIONS (IF PRESENT, MUST IMMEDIATELY FOLLOW THE VOWEL)

Value	Stress Assignment	Value	Stress Assignment
0	No stress	3	Tertiary stress
1	Primary stress	.	(Etc.)
2	Secondary Stress	:	

Table 1.1 ARPAbet symbols for representing within a computer the phoneme-like units of English

(from "Trends in Speech Recognition", p-127, see (1))

This dissertation is mainly based on the project of the development of the phoneme-based word recognition system. The final system is named SWIS (See What I Say), because extensive graphical representations have been employed during the system development and evaluation. The purpose of developing SWIS is twofold.

Firstly, it tries to answer a basic question, that is, is it feasible to design an automatic speech recognition system based on the model proposed in APT. Obviously, any sufficiently verifiable solution to such a problem is indistinguishable from a working system. That is the main motivation for developing SWIS.

Secondly, an automatic recognition system such as SWIS will enable us to test several key concepts introduced in APT on speech in a quantative fashion (i.e. using large data corpus collected from many users) with high consistency (i.e., no manual intervention). SWIS is not intended to be perfect but intended to explore the problems which might only be discovered from developing a working system.

This dissertation will discuss SWIS in great detail. Not only is the functionality and the design philosophy of each system component discussed but also the detailed stepwise algorithms to implement these components are presented. Because this study is a part of a large project involving teamwork, it is very important for every member in the team to know exactly how SWIS is implemented. That is why the stepwise algorithm form is used in the following chapters to describe SWIS so that this paper can serve as a comprehensive system documentation. Other readers of this paper may skip these algorithm sections if they are interested only in the logic behind the designing of these

components and the performance on the testing databases which will be described in this chapter.

In Chapter 2, we will review the basic concepts of APT and discuss the technology required to design a specific recognition system based on a generic model proposed in APT. The signal processing and feature representation are described in Chapter 3. All of the system components considered by the author as "front-end," such as broad phonetic classification and sensory path generation (including formant tracking) are described in Chapter 4. In Chapter 5, perceptual path generation, parsing, and phonetic dictionary-based lexical accessing are described in the context of SWIS recognition process. The experimental results of SWIS recognition procedure at both phoneme level and word level are presented in Chapter 6. The notes on SWIS system implementation described in Chapter 7 are written mainly for those who will use and modify SWIS in their future research.

In the remainder of this Chapter, we provide a brief background discussion on ASR and an overview on the problem areas associated with the current research on the feature-based recognition systems so that readers may have a better understanding of what we are trying to accomplish in this study in a perspective context. The research goals and databases used are discussed at the end of this chapter.

1.2 BACKGROUND

An ASR system generally falls into one or any combination of the following four categories: (1) speaker-dependent Isolated-Word Recognition (IWR); (2) speaker-independent IWR; (3) speaker-dependent Continuous-Speech Recognition (CSR); and (4) speaker-independent CSR.

In the following three sections we will discuss the hardware and technology requirements and the performance in each category. The performance of an ASR system is usually evaluated in terms of vocabulary size, response time, recognition accuracy, training procedure and the complexity of the language to be recognized.

1.2.1 Isolated-Word Recognition

As commonly defined, isolated-word speech is spoken with distinct pauses between words or items. An item is a short phrase stored as a word; the phrase must be spoken without pause and is simply treated as a single "word". Thus, "load disk A" might be entered into an IWR system as a single word. As such, "load", "disk", and "A" are not available as single words.

Today, commercial IWR systems cost from $500 to $8000. These systems usually have an acoustically distinct vocabulary of 10 to 1000 items, generally require clear pauses between items, and need "tuning" for each user (the details of the system tuning are discussed in section 1.2.3). An item is usually a single word or a short phrase like "load disk A" but is generally restricted to a time interval of no longer than 2 seconds.

Algorithms for developing limited-vocabulary, speaker-dependent IWR systems are well understood. Because each item is treated as a unit, recognition simply calls for comparing the parameters of the input item to **EVERY** prestored item-template. The item with a stored template that best matches the input is selected as the intended item.

The problems of speaking rates have been to a large extent solved by a variety of time-alignment procedures. The most successful is the

Dynamic Programming Warping (DPW) introduced in Japan in 1971 (26-27). A typical IWR system requires only one or more microprocessors, dozens of digital filters, and 128-512K of memory. All of those components are available on the current market at fairly low prices. With the increasingly mature VLSI technology, an IWR on a single-chip is expected in the near future.

1.2.2 Continuous-Speech Recognition

First, we will make a distinction between connected speech and continuous speech by using the definitions by Meisel (28) for these two confusing terms in the next two paragraphs.

"Connected speech is isolated-word speech without a clear pause. For connected speech to work well, each word or word-equivalent should be stressed. Also it is usually required that there be no significant coarticulation between adjacent words."

"Continuous speech is speech spoken without unnatural pauses, with a natural stress pattern, and without restrictions to avoid coarticulation. This is best defined as speech the way a person might speak if he were not attempting to think about the way he is speaking — natural speech."

In addition, we like to add to the above definition that continuous speech consists of continuously spoken sentences not phrases. As such, we do not consider any system which recognizes continuousely spoken English digit sequences or other similar sequences as a CSR system.

Before the start of this project, there had been only two reported systems that met the above definition for continuous speech (HARPY and HEARSAY-II) with acceptable recognition accuracy (usually 90% or

higher). Both were developed at Carnegie-Mellon University as a part of DARPA SUR project (29-30). However, neither operates in real time so far; HARPY would require a machine with 28 MIPS* and HEARSAY-II would require a machine with 85 MIPS to do so.

During the time SWIS was being developed, the research on CSR has witnessed the significant technology advancement marked by the fact that scientists now can build a microcomputer-based CSR machine which recognizes truely continuously spoken sentences with a demonstrated sentence recognition accuracy of 95% (by Speech System Inc., CA). Incidentally, this system is also phoneme-based. However, SSI's machine recognized only the sentences completely prespecified in a very simple recognition task.

There are several connected speech recognition systems available today. They usually require a brief pause (generally between 200 ms. and 2 seconds) between words and recognize a very restricted language with a vocabulary of 100 to 1,000 words. The only exception is the IBM prototype talkwriter that recognizes standard English with a vocabulary of 5,000 words (31).

Comparing IBM's "great statistical engine" ($10 million development cost) with SSI's "Phonetic Engine"** in terms of performance and user acceptability we see a classical debate on the question of whether CSR is really needed? In the case of the IBM machine, it recognizes any speech if one is willing to speak "slowly." In the case of the SSI machine, it recognizes speech no matter how it is said, but one must

*MIPS: Million Instructions Per Second executed by a computer.
** Phonetic Engine is the registered trademark of SSI.

tell the machine in advance every possible sentence that one might utter. We are not going to get into detailed discussion on this issue because it is beyond the scope of this paper.

Algorithms for developing today's CSR are much more complicated than ones for IWR and are often dependent on interdisciplinary study across several research areas such as psychology, phonetics, linguistics, and mathematics. Several representative models proposed to study CSR are the statistical models in the IBM System (31), the Markov models in the BBN system (20), the Syntax-Directed DPW model in the Bell system (32), LITHAN at Kyoto University (33-35), the EAI (Empirical Artificial Intelligence) model in the SSI system (36), and the APT model on which SWIS was originally developed. Recently, Colla et al. proposed a single network using a diphone-based language model, in which spectral sequences of diphones were used (37).

It is very difficult to predict which technology behind these systems will become the best candidate for solving the problems in CSR. But it is safe to say at this writing that none of these systems will "grow-up" in the near future (say 5 to 10 years) to meet the ultimate goal of CSR — a second generation speech recognizer that recognizes natural speech by anyone speaking in general English.

It is my belief that in order to develop second generation speech recognizers two objectives need to be achieved: (1) the relationship between the acoustic properties of the speech signal and the speaker-intended phonemes have to be fully understood; (2) a programmable knowledge base is needed, which contains both large statistical data from real speech and so-called rules of thumb derived from human experts

in the areas of acoustics, phonetics, linguistics and knowledge-engineering. Furthermore, the realization of the above-described knowledge bases will depend on computers with speeds of 100,000 MIPS in order to achieve real time response. These computers will require substantial advances in both VLSI and supercomputing technologies.

1.2.3 Speaker Independence

An IWR or CSR system is speaker-independent if it recognizes speech selected randomly from speakers of most dialects in that language. Research at Bell Laboratories shows that up to 12 different speech patterns are needed to represent the way a single word might be pronounced in the United States (38). Because there is a very large population with many dialects, it is very difficult to verify that a system is truly speaker-independent. Many systems are claimed to be speaker-independent after testing with small populations, say 10 to 100 speakers. Even so, only a few systems on today's market are speaker-independent in this sense, and they are limited to the recognition of a simple vocabulary like the numbers 0 through 10 or simple words like yes and no. In the remaining portion of this report we will use "DIGIT" to refer to the English digit vocabulary just mentioned.

For the above reasons, the term "System Tuning/Training" has been developed to distinguish speaker-dependent from speaker-independent systems. Thus, for a speaker-dependent system a new speaker generally has to speak EVERY word in the vocabulary repeatedly in order to make the system understand his or her voice. On the other hand, speaker-independent systems are capable of recognizing speech from an unknown

speaker (i.e., the system does not possess any knowledge of the speaker prior to the recognition).

The training procedure for a large vocabulary like one used in the IBM system may take one hour or two, in which time a new user reads a carefully designed script a few times. The purpose of the training procedure is for systems to correlate the acoustic properties of signals with the known words and then construct speaker-dependent templates for every vocabulary word.

In phoneme-based systems, the training procedure for building a speaker model may takes even more time than IBM's counterpart. In this case, the length of time and the cost of the new speaker enrollment procedure enable the system to correlate the acoustic properties of signals with the known phonetic spelling for many carefully selected phonetic-balanced sentences. For example, find onset time, burst frequency and transient properties of /t/ for all the possible sentence contexts.

Here, the key issues are (1) to discover features that do not vary from speaker to speaker, which is one of the main goals set in APT; if it is ever realized then the training procedure will be completely eliminated; (2) to design the optimal training procecdure to reduce the training time to no more than 20 minutes, assuming that it is acceptable by the general public; and (3) to reduce the cost dramatically by using advanced DSP chips.

The rationality of the last issue relies on the fact that if a talkwriter of tomorrow is to sell for below $200, as a typewriter does

psycholinguistics, there is evidence which supports the theory that the auditory input is stored in a preperceptual memory and the output of the process that operates on the data in this memory is a syllable representation (43-44). In spite of the fact that the theory is difficult to prove, many psycholinguistic theorists believe that the basic unit of speech perception is the syllable (45).

Besides the influence from psychologists, the major motivations of designers of ASR for changing the recognition unit from word to syllable are: (1) to have a fixed number of reference templates that do not increase with the vocabulary size, (2) to reduce time to access the stored templates, and (3) to study continuous speech.

With the syllable template, many languages can be represented in a very small number of sets. For example, one study shows that Japanese can be represented in a hundred syllables (12). Even for English which has about ten thousand syllables, the improvement with syllable representation on both memory requirement and access time, when compared with a word template, is substantial if an adult English vocabulary (about 100,000 words) is used (for example, in the second generation feature-based talkwriter).

As far as English is concerned, two major problems arise from the syllable representation. First, for many syllables in unstressed mode, like the *second* syllable /V+AX+N/ (see Table 1.1 for ARPAbet notation) in the word SEVEN, their acoustic realizations vary in different intersyllabic contexts. Secondly, the criteria of syllable segmentation depend upon detailed knowledge of articulatory-acoustic

relations, which cannot be easily represented in computer understandable forms.

1.3.1.3 Demisyllables

Demisyllables are defined as half-syllable units which occur in strictly constrained initial-final pairs. The initial demisyllable is generally made quite short, extending just beyond the initial CV transition. In this way, any influence of postvocalic consonants (especially nasals) is largely confined to the final demisyllable. As an example, the word SEVEN can be represented in a demisyllable sequence: /S+EH/+/EH+V/+V+AX/+AX+N/.

One of the advantages of the demisyllable unit over the whole syllable is large reduction in the size of the reference inventory. One study (46) shows that a reduction by about a factor of five can be obtained for some medium-sized vocabularies (100 to 300 words). The reason is that the same demisyllable can be used for more word compositions. For instance, the demisyllable /S+EH/ can be used to compose the words SEVEN, SEPTEMBER, SEX and etc.

Another advantage is that the demisyllable segmentation is more efficient than the syllabic segmentation in terms of recognition accuracy; this is especially true for a large vocabulary consisting of 1,000 words or more (47). As far as the segmentation errors are concerned, detecting the syllabic segment boundaries is just as difficult as it is for detecting demisyllables. But with the demisyllable representation, an average at least twice the number of recognition units are generated as the ones in syllable representation for the same dictionary. As such, the same number of segmentation

errors would take a relatively smaller percentage in demisyllable mode compared with syllable mode.

In addition to the same shortcomings as with the syllable unit discussed in the previous section, the demisyllable scheme requires an earlier decision on cutting an uncertain CVC syllable into the initial demisyllable and the final demisyllable. In case of diphthongs such an earlier cut may lead to the later recognition of mistakes in the two adjacent syllables with a final vowel followed by an initial vowel (for example, the phrase "buy ink").

1.3.1.4 Phoneme

Definitions for the "phoneme" can be found in many textbooks on phonetics and linguistics and in many scientific journals (23,48-49). For our purpose, the phoneme is the shortest segment that makes a significant difference between utterances. These segments are linguistic units and have their bases in speech as it is spoken and perceived.

The advantages of selecting the phoneme as the recognition unit are well known. First, all of the syllables and words in any language can be represented by a relatively small set of phonemes. For example, it is generally considered that English has about 25 consonants and 12 vowels for a total of 37 phonemes (23). The second advantage is that the phoneme maps in a more direct way to lexicon entries which are usually phonemic-oriented.

The phoneme-based recognition system (PRS) has been the major target aimed at by the ASR scientific community in this decade, because once realized it could offer a host of attraction (20,50). In PRS

training procedures, memory requirements and accessing time can be greatly improved even if it is speaker-dependent. First, a text covering all the phonemes could be read by a user in a matter of minutes and the machine could be trained for any size vocabulary. Second, it requires much less memory storage for phoneme templates. Thus, for the same memory configurations mentioned in section 1.3.1.1 the storage of 150 KB can be used by 2 talkers for 20,000 words or more, rather than by 1 talker for 500 words (assuming 46 phonemes or less in the underlying language).

Because our main interests are in the phoneme-based system, we will not discuss the problems with phoneme-template in this section. Instead, we will devote the next three sections entirely to examining the problems with segmentation, feature extraction and cue identification in the context of phoneme-based systems.

1.3.2 Segmentation

Ideally, one would like to classify directly the continuous speech signal into segments that correspond to **detailed** phonetic events. However, automating the task with high accuracy is very difficult due to the high degree of acoustic variability in the speech signal, although some human experts can manually perform the task with 85% accuracy (14).

With the current limited understanding of acoustic-phonetic correlates, as Klatt points out (51) the errors in segmentation are inevitable no matter how the task is accomplished. The best strategies have been in the passive mode: (1) to avoid it (don't do it); (2) to defer it if it cannot be avoided (do it as late as possible); (3) to

make the broad phonetic classification (if it cannot be deferred) at an earlier stage and then refine it later.

Leung and Zue (52) use six phonetic classes as initial segmentation decisions in a system for automatic alignment of phonetic transcript with continuous speech. The six phonetic classes are vowel-like sonorant, obstruent, voiced-obstruent, silence, nasals and sonorant-energy dip. Even with such a broad classification, segmentation will still run into the problem with ambiguous class boundaries. For example, it is often difficult to locate the boundary between an unstressed vowel segment and the following nasal segment, as in case of word "SEVEN".

The vector quantization techniques applied in the past to fixed-length segmentation (22,53) have been recently extended to Dynamic Segmentation (DS) so that the resulting segments are able to represent meaningful acoustic-phonetic units (54). In DS models each token is modeled by a codebook sequence containing as many as the number of segments in the corresponding handcrafted network. A standard DPW algorithm is then used to pair a segment in the unknown with a segment in the network, so that the total quantization distortion is mininal.

The problems with the DS recognition system are: (1) it is more difficult to train (since handmarked segments are required); (2) more computation is needed during recognition. In addition, the vector quantization techniques assume that all the segments have the same weight, regardless of their semantic relevance within a word. This assumption directly contradicts the fact that humans can perceive the

meaning of a spoken word with inserted or omitted phoneme segments because of their knowledge of language.

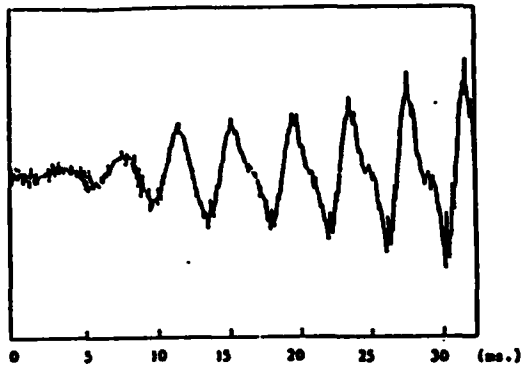
1.3.3 Feature Extraction

There is extensive justification for characterizing speech as a finite set of simultaneous features associated with an acoustic signal representing a phoneme (55-59). After many years of basic research in the acoustic theory of speech production and auditory physiology and psychophysics, we can now identify acoustic features that are important cues for the perception of almost all the segmental phonemes (10,49). The questions remaining to be answered are how many acoustic features are needed in a feature-based system and how to extract these acoustic features or parameters in terms of accuracy, efficiency and consistency.

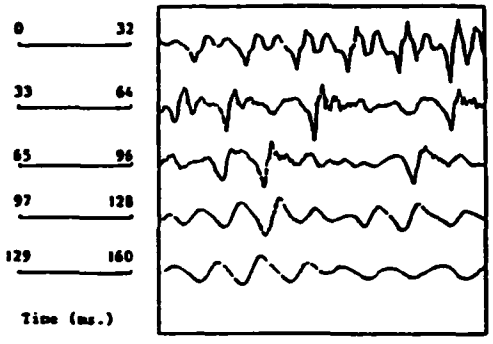
Different recognition systems often chose different numbers of features dependent on the underlying system structures. In an extreme case, the English letter recognition system FEATURE (60) recently developed at Carnegie-Mellon University depends on as many as 50 acoustic features (61-62).

Some features such as zero-crossing rates can be easily extracted from the acoustic signal since their underlying measurements can be described in analytical forms. But extractions of others can be extremely difficult. The two best-known examples are Pitch Extraction and Formant Estimation. Continuously active research in these two areas has been recorded in a large body of literature (63-71).

The problems in pitch extraction are mainly caused by the transition between periodic and aperiodic speech signals. Figure 1.1 (a) shows a periodic waveform segment. Note that the pitch pulses in

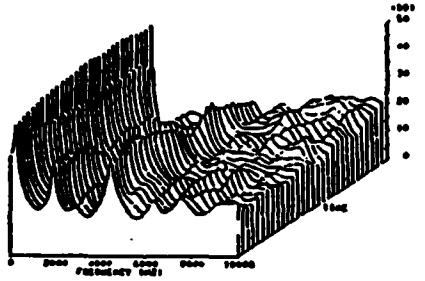


(a) the transition between /Z/ and /IH/ in the utterance "ZERO".

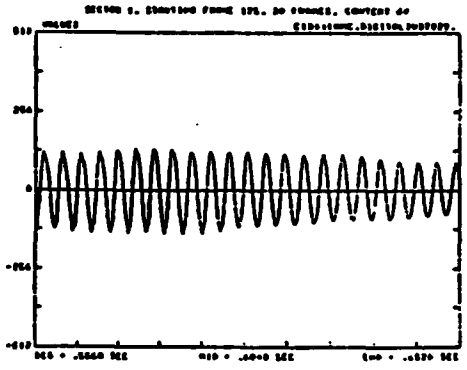


(b) the continuous transitions from /V/ to /AX/ and then to /N/ in the utterance "SEVEN".

Figure 1.1 The sample plots of the speech waveforms showing the irregularity of "pulse chains".



(a) the LPA spectrums are computed from a /N/ segment in the utterance "SEVEN".



(b) the waveform from which the spectrums shown in (a) are derived.

Figure 1.2 Illustration of the problem "peak-mergin" in formant estimation.

the beginning of the waveform are very weak and therefore locating the first pitch pulse is a non-trivial problem. Another problem is that not every pitch pulse shows up in a periodic waveform. This can be illustrated in the waveform of a sonorant segment shown in Figure 1.1 (b). A possible explanation for missing pulses is that F1 is so close to F0 (fundamental frequency) due to the nasalization, that the pulse chain has been corrupted.

As commonly defined, formants are the resonant frequencies of the vocal tract, which in turn must be estimated from "prominent" peaks in short-time spectral "envelopes". The problem first arises from the ambiguous definitions of both "prominent" and "envelope". How can one obtain such an envelope and what is an appropriate threshold for a prominent peak (e.g. an 8 db. local maximum peak)? The second problem is how to handle missing formants. An example of a set of linear prediction spectrums is shown in Figure 1.2 (a). Most formant tracking algorithms would consider the first (strongest) spectral peak as fundamental frequency and pick the next three peaks as F1, F2 and F3 (first, second and third formant). However, if the formant tracker knows that the current segment (the waveform corresponding to the spectrums is shown in Figure 1.2 (b)) is nasal-like and the speaker's pitch is about 250 Hz., it will compute formants based on completely different criteria. In this case, the upper edge of the first spectral peak will be taken as F1 because the first formant for the nasals is very low, typically below 300 Hz. Here, the key to successful feature extraction of algorithms is to incorporate other sources of knowledge to overcome the inherent uncertainty in "front-end" analysis.

1.3.4 Cues to Phonetic Identities

Acoustic cues to phonetic identities, also called acoustic invariance of phonemes, have been studied by the speech community for a long time. The discovery of the /D/ locus (49) is the classic example for demonstrating that there are measurable acoustic cues for some phonemes in certain phonetic environments. Ideally, we wish to find the acoustic cues for every phoneme in the underlying language, which are invariant across not only different syllable environments but also across different speakers. For example, we know that the phoneme /W/ has the lowest F1 relative to other phonemes. But we cannot simply scan an arbitrary F1 contour and label the segment of the lowest valley as the phoneme /W/. This is because (1) the acoustic feature is not one-to-one corresponding to a phoneme and (2) the phonemic information overlaps more than one segment (6). In other words, the identity of a phonetic segment depends on the identities of its surrounding segments. It is well-known that phonetic recognizers based upon this kind of mutual dependency are difficult to implement.

Whether there exists a set of acoustic invariant cues for each phoneme is not yet known. Even if we had such a set of cues today, we would still have to solve problems with the omitted and inserted phonemes which often occur in fast speech and continuous speech. In the case of phoneme omission, the corresponding acoustic cues are never realized in the acoustic signal. For example, the phoneme /T/ in the word "EIGHT" may be intended by the speaker though he may not actually produce that sound. Sometimes, an extra phonetic segment has been

inserted due to the particular way the speaker pronounced it. In this case, the inserted acoustic cues have to be ignored.

In continuous speech, the same phoneme may be adjacent to the next word other (e.g. in the phrase "Bus Stop"). Then either the same phonetic segment must be partitioned into the two subsegments so that the underlying acoustic cues are invoked twice, or an extra phoneme has to be inserted at a later stage based on the high level knowledge.

From the system design viewpoint, identifying a phonetic sequence correctly is just half way to the final goal - the recognition of speaker-intended words. Those phonetically similar words like "ANN" vs. "AND" and other short words like "THE" and "A" may only be recognized by the context.

1.4 LITERATURE REVIEW

The research in the feature-based recognition system using syllables or smaller units has been quite limited so far in the United States although some large efforts have been undertaken in Japan (72) and Europe (50,73). Most of these studies are limited to the recognition of a few phonemes like stop consonants (74-75) or a few vowels (24-25). Makhoul and Schwartz (20) developed a recognition system based on a Hidden Markov Model (HMM) representation. With such a model they are able to study the context-dependent properties of all the English phonemes. None of these systems address the issue of developing context-independent phonetic targets (partially because many still doubt the existence of such targets).

The formant-based recognition strategies similar to some core concepts created in APT have long been studied by the speech recognition

community. As early as 1952, for example, Davis and his associates (76) were able to construct a speaker-dependent English digits recognizer by using the formant-based word templates from F1 and F2 parameters. In the late 1960s, it was noted by Liberman (49) that place of articulation for stops can be determined from formant motion into the adjacent vowels. In the 1970s, formant-based speech analysis began to cover more classes of phonemes than just vowels, such as the work of Weinstein (77) which involved spotting velar stops from the proximity of F2 and F3, and diphthongs from formant motions. In 1976, Woods (78) reported that articulation of all the plosives and nasals can be classified on the basis of formant motions. In addition, Kameny suggested (79) that the identities of semivowel /W, Y/ and liquids /R, L/ can be found by studying formant behavior.

As we will see in the next chapter, the development of APT really goes beyond the scope of all the formant-based speech analysis work mentioned above. In essence, APT tries to create a broad framework wherein all the English phonemes can be represented by a set of formant-related parameters, in a uniform fashion. However, we must point out that those formant-related parameters proposed in APT are far more complicated than commonly understood formants.

1.5 THE RESEARCH PLANS AND GOALS

One long-term goal of research in feature-based phonemic recognition is to determine the characteristics of the phonetic targets in terms of acoustic features. The reason is simple. The value of a finite set of acoustic features or parameters extracted from a phonetic segment will activate **one and only one phoneme** target, no matter by whom

the underlying phoneme was voiced and how it was pronounced in terms of phonetic environment. However, the discovery of such a speaker-independent target for every phoneme in English will take many years of research. As we pointed out in the previous sections, today we simply do not fully understand the phonetic identity of speech sounds and their acoustic correlates. Until complete understanding has been achieved we should carefully control the complexity of the language and the number of talkers involved in the research. In this project we limit the research to a simple vocabulary, hopefully gaining better insight into a few English phonetic targets in a restricted phonetic environment.

In the next three sections, we first lay out our research plans and the goals in detail. Then the acquisition of acoustic data will be described. Last, we will discuss the limitation of our study and the resources involved in the project.

1.5.1 The Project Scope

This project is to develop a prototype speaker-independent recognition system named SWIS for a selected set of phonemes derived from all the stressed syllables in the vocabulary DIGIT as defined in section 1.2.3. Table 1.2 lists all the words in DIGIT and the phonemes selected for this project are listed in Table 1.3.

Even a small vocabulary such as DIGIT may contain many phonemes, as noted in Table 1.3. It is difficult to study all these phonemes in one project. But, syllable contexts will be greatly reduced if we focus our attention on stressed syllables in DIGIT. Furthermore, the acoustic cues for phonetic segments around stressed syllables are usually far more reliable than they are around unstressed syllables, according to

TABLE 1.2 DIGIT VOCABULARY

WORD	ARPAbet
ZERO	Z+IH+R+OW
ONE	W+AH+N
TWO	T+UW
THREE	TH+R+IY
FOUR	F+UH+R
FIVE	F+AY+V
SIX	S+IH+K+S
SEVEN	S+EH+V+AX+N
EIGHT	EY+T
NINE	N+AY+N
TEN	T+EH+N

TABLE 1.3 Phonemes Studied in SWIS

CLASS	PHONEMES	WORD CONTEXT
VOWEL	IH	ZERO, SIX
	AH	ONE
	UW	TWO
	IY	THREE
	UH	FOUR
	EH	SEVEN, TEN
DIPHTHONG	AY	FIVE, NINE
	EY	EIGHT
CONSONANT	W	ONE
	R	ZERO, THREE, FOUR
	N	ONE, NINE, TEN
	S	SIX, SEVEN
	T	TWO, TEN

the study of Cutler and Foss (80). In addition, the phonetic segments within the stressed syllables of DIGIT provide powerful constraints for lexical access because there are only eleven words in DIGIT.

The idea of recognizing stressed syllables can be generalized to work for large vocabularies. For example, a recent study conducted by Huttenlocher and Zue (81) indicated that 1/3 of the words in a 20,000 word lexicon (the Merriam Pocket Dictionary) can be uniquely specified in broad phonetic classes. For example, the following activation sequence of phonetic-class targets:

[Consonant][Consonant][Liquid or Glide][Vowel][Nasal][Stop]

will limit the 20,000 word lexicon mentioned above to two possible entries (one is "SPLINT"). According to their study, over half of the lexical items belong to equivalence classes of size 5 or less. Note that no detailed phonetic feature is used to identify each member of the VOWEL class in the above-described study. It is our hope that if we can recognize stressed syllables we can further reduce the number of possible sequences that satisfy broad phonetic class descriptions.

The main strategy employed in the SWIS development is to map directly acoustical parameters to phonetic events by means of a phonetic-encoded auditory-perceptual map. Thus, any spoken word can be "seen" on such a map as a path with many commonly familiar properties such as origin, destination, direction, travel distance and time, and so on.

The purpose in developing such a system is to verify the concept and the feasibility of the theory. We will design new algorithms to obtain formant information from vocalic segments of natural speech, so that the stressed syllables can be represented in an auditory perceptual space as continuous paths upon which a rule-based phonetic parser can be applied, to identify phonetic segments with high accuracy. Although a side effect of recognition of the phonemes in the stressed syllables of DIGIT is the recognition of eleven English digits, the goal of this project is to establish a framework for study of phoneme-oriented IWR based on the current development of APT.

1.5.2 Acoustic Data Acquisition and the Database Used

The acquisition of acoustic data was carried out in the laboratories of the Research Department at the Central Institute for the Deaf in St. Louis (CID) where the SWIS was developed. All the subjects who participated in the recording were either employees of CID or graduate students who work part time at CID.

Three data corpuses, namely, TRAIN, TEST1 and TEST2, were used in this project. The information on each data corpus is listed in Table 1.4. The data corpuses TRAIN and TEST1 are used in the development of SWIS and the initial test, while TEST2 is used in the final test to evaluate SWIS performance. The talkers are adults between ages 25 and 60, recorded in the anechoic chamber using the high quality B&K 4179/2660 microphone. The speech level is monitored at 60-65 dbA at the microphone. The recordings are made directly on a Sony PCM-501ES digital audio processor with 16 bit amplitude encoding at 44.1 kHz sampling rate.

TABLE 1.4 DATA CORPUSES USED

DATA CORPUS NAME	TALKERS		WORDS	TOKENS
	Male	Female		
TRAIN	1	1	11	44
TEST1	1	1	11	44
TEST2	11	10	11	231

Note: The talkers for TRAIN and TEST1 are recorded twice for each of 11 digits to produce 44 tokens.

For the corpuses TRAIN and TEST1, the audio signals are monitored through a Tektronix 7313 oscilloscope for excessive peak clipping or underusage of dynamic range. The speech signals are also edited on an inhouse digital signal processing system called PARAPET, run on a Data General Eclipse 200 through a playback unit for detecting token boundaries. The resulting signal corresponding to each token is then redigitized at 20 KHz. on PARAPET with 12-bit quantization to form an Eclipse disk file. Then the files are transferred through a 9" industry standard tape to a micro Vax-II for subsequent processing.

For the corpus TEST2, the audio signals are recorded and then edited on EDITS (an in-house digital signal processing system) run on a PDP-11 for the same purpose as mentioned above. The resulting signal corresponding to each token is then redigitized at 20 kHz with 16-bit quantization to form a PDP-11 disk file. Then these PDP-11 RMS files

are transferred to a micro Vax-II by running KERMIT on both computers to generate Vax VMS files.

The talkers for TEST2 are actually recorded twice for each of 11 digits, that is, a total of 462 utterances are recorded. Because our disk space on VAX computers is very limited, only one recording per talker per word is chosen to form the data of corpus TEST2. The decision of which recording (one out of two) is chosen is made on the basis of correct pronunciation judged by listening to the playbacks of each recording. This prescreening process is done carefully by an experienced research staff at CID to minimize pronunciation errors. Among the finally selected 231 tokens only one is considered as mispronounced. It, spoken by a female subject, sounds like "SEX" instead of "SIX" when played back.

The talkers speak with the same midwestern accent, except for several in TEST2 who speak with a southern accent. Therefore, the SWIS was exclusively "trained" by the midwestern talkers, but tested by talkers with different accents. The impact of this dialect difference on SWIS performance will be discussed in Chapter 6.

1.5.3 Limitation and Research Source

One of the major limitations of this project was that speech editing and playback facilities are housed on one computer and the system development for SWIS is done on another. At the beginning of this project, all the digital signal processings such as editing and playback were performed on an Eclipse minicomputer. Because the programming environment on Eclipse is relatively limited, it was decided to develop the entire SWIS system on a micro VAX-II running the VMS

operating system. However there was no direct communication link between our DATA GENERAL Eclipse computer and the Digital Equipment Corporation VAX computer. All the waveform files generated on Eclipse had to be transferred through a 9 track 3/4" tape to VAX computer by running an inhouse data format conversion program.

Later, the EDITS was used to create waveform files on a PDP-11 because at that time we were able to run KERMIT successfully on both micro VAX-II and PDP-11 through a point-to-point link at 9600 baud rate. Although transferring a large speech database of megabytes magnitude using KERMIT is not ideal, the operation is much simpler than one involved in using a 9 track tape as intermediate media.

The software development for SWIS could be done on a PDP-11, taking advantage of using the playback unit integrated with the computer. There were two factors responsible for the selection of the micro VAX-II rather than the PDP-11. First, we could not allocate a PDP-11 at CID exclusively for this project in spite of the availability of accessing several PDP-11s on an hourly basis. In contrast, we were able to use a micro VAX-II almost exclusively for the SWIS project. Second, a commercially available signal processing package ILS (Interactive Laboratory System, Signal Technology Inc., CA) has been used extensively at CID in several other projects closely related to SWIS. Furthermore, we had only the VMS version of ILS to run on a micro VAX-II but not on the PDP-11. Thus, it was desirable to have all the software packages including SWIS run under the same operating system so that many speech analysis parameters could be created in a consistent manner.

The VMS operating system on the micro VAX-II used at CID provides a very flexible programming environment for SWIS development. However, there is no playback unit on our VAXs and moreover no software support even if we did have a playback unit. Therefore, there is no convenient way to listen to the playback of a particular segment in an utterance and to see the related features at the same time. This inadequacy makes the process of manually labeling broad phonetic classes and detailed phonetic segments not only difficult but also time consuming. For example, the boundaries between nasal segments and vowel segments are very difficult to detect just by looking at short-term spectra and other features extracted from the segments.

Another major limitation of the computer resource for SWIS is that we have very limited disk space on the micro VAX-II on which SWIS is developed. It has only 71 MB of disk space. The VMS operating system and ILS take 40% of the total disk space. In order to accommodate other users of the system, the actual disk space allocated for SWIS project holds just about 1 minute of speech data and related parameter files besides SWIS software.

With such a storage configuration, SWIS can be tested against all the speech data files in both TRAIN and TEST1 at one run when they are all stored in the same computer. If we had a large train database, say about 5 minutes of speech, we would have to break each test run into 5 operations, that is, run a subset of the database at one time, delete it once finished, copy another subset of the database from either a tape or another computer system through a network, and then run the subset.

Furthermore, every time a parameter or a threshold in SWIS is changed, all the test data files have to be rerun to check whether these changes work for all the test database. It is our past experience that many speech recognition algorithms often require hundreds of test runs to reach satisfactory performance. Anticipating the limited disk storage for this project, the size of the train database (TRAIN and TEST1) was kept so small that the ratio of the size of the train database to the size of the final test database (TEST2) was even less than 1. Comparing with several milestone systems of the past (HEARSAY-II and HARPY), all had much larger values than this train-to-test ratio. It is our view that any claimed speaker-independent recognition system should have a ratio of less than 1 when the recognition scores are computed from whatever test databases are created by new talkers.

Because of the relatively small size of the train database used in the SWIS project, the process of software development for SWIS is quite efficient as far as the test runs are concerned. The disadvantage of using such a small train database is that many important parameters and thresholds used in SWIS are tuned to fit this data set (created from 4 talkers). It is inevitable for SWIS to fail to recognize some important phonetic segments when it is tested against a large-size data set, say, created from 20 or more talkers. The consequence of using such a small database on the performance of SWIS will be described in Chapter 6.

2. THE SYSTEM CONCEPT AND SWIS OVERVIEW

2.1 INTRODUCTION

In this Chapter we discuss the basic system concepts used in SWIS and the system working assumptions on which SWIS is designed. The major components of SWIS are also described in conjunction with the specific goals of each component.

As we stated at the very beginning of this paper the motivations for building SWIS and the main ideas for how to build SWIS are all inspired from studying several important concepts created by Miller in APT (the Auditory-Perceptual Theory) and working closely with Miller at CID. In fact, the SWIS was originally proposed as a part of a large scale investigation titled "The Auditory-Perceptual Theory of Phonetic Recognition".

It is very important for the readers of this paper to understand the basic ideas and some core concepts created in APT. However, at this writing, the major manuscript of this theory (APT) has not been published, although it has been presented at several technical conferences in the past. For this reason and for the sake of completeness, we include in this paper as Appendix 10.1 the important elements of a description of APT. We strongly suggest that the reader read Appendix 10.1 before continuing to section 2.2.

Before readers proceed to the next section, we would like to make some general comments regarding the background of the development of APT. Because the main goals of the investigation of APT were to try to understand how human auditory-perceptual mechanism works, the tone and

the terms used in APT are quite different from what one might read in typical scientific literature for automatic speech recognition. In other words, APT is more concerned with the problems associated with speech perception than with those of speech recognition. In addition, the backgrounds of scientists in the area of speech perception are traditionally different from those of scientists in the area of speech recognition. This difference often introduces communication gaps between the two groups. So, readers with engineering backgrounds need to pay more attention to Appendix 10.1 where the main ideas of APT are described by Miller himself rather than a third party.

The importance of close cooperation between the two groups of scientific researchers mentioned in the previous paragraph has been increasingly recognized by both groups. Great efforts were made during the development of SWIS to materialize as many ideas of APT as possible. SWIS can be considered as a joint product of both groups, in the sense that it tries to apply the knowledge learned from the studies in speech perception to solve the problems in speech recognition. As such, we need to compare some new concepts, terms and problems addressed in APT with similar but more well-known counterparts studied in the past. This comparative review of APT is given in the next section. After reviewing the basic concepts introduced in APT, we will discuss the basic scheme used in SWIS. The overall system organization of SWIS is also presented in the next section. Each component of SWIS is briefly discussed in terms of its functions and goals. At this point, we strongly suggest that the reader read Appendix 10.1 and then come back to proceed to the next section.

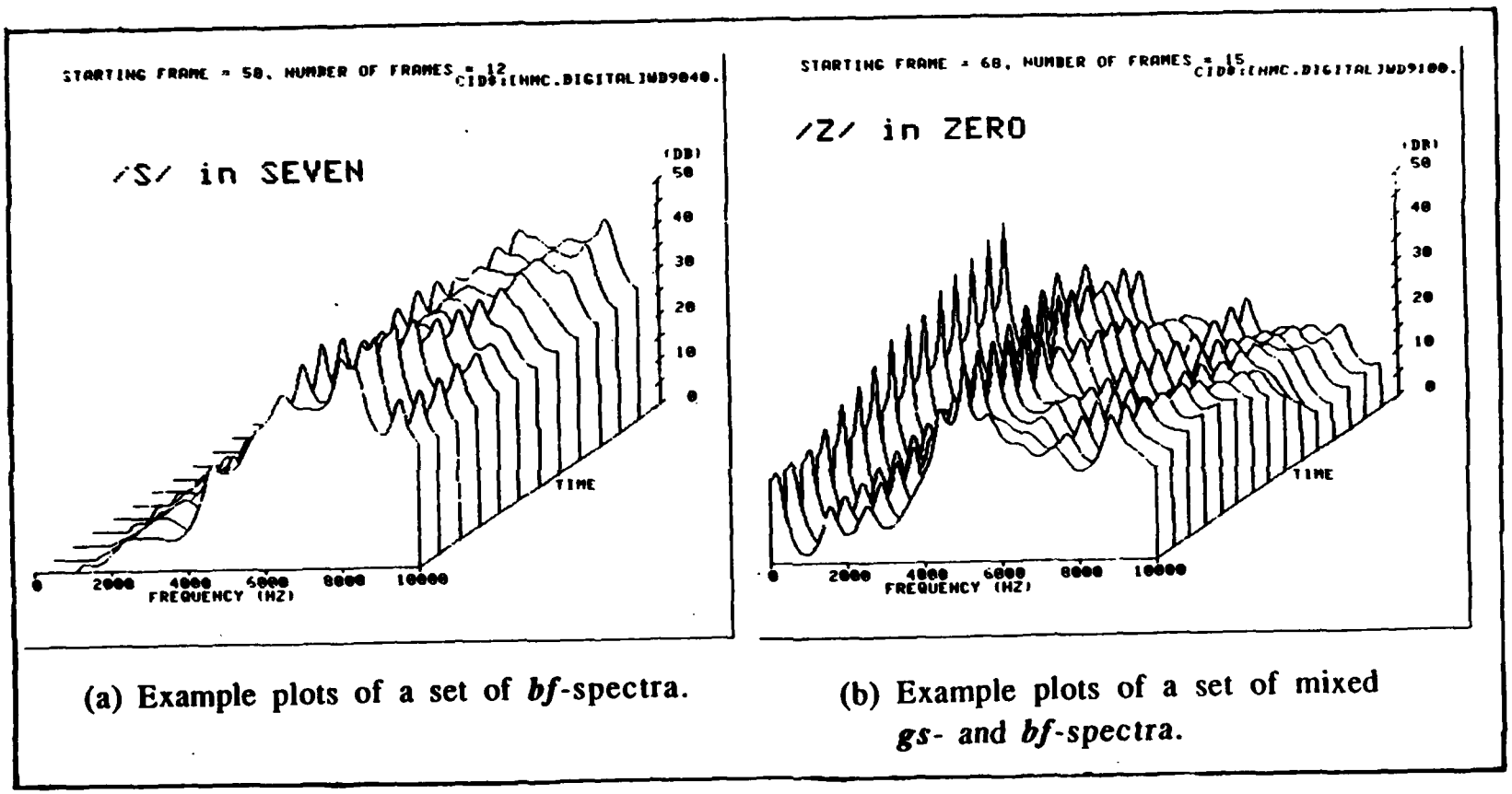
2.2 SYSTEM WORKING ASSUMPTIONS AND SWIS OVERVIEW

2.2.1 Comparative Discussion

Because there are many new terms introduced in APT as we have seen in the previous section, we would like to highlight several key concepts relevant to this project through a comparative discussion in the more familiar terms found in the literature of automatic speech recognition. By doing so, we can put all these new terms and concepts in perspective so that the reader can follow our basic working assumption under which the goals of SWIS are pursued.

To simplify our comparative discussion below, we assume that acoustic signal from natural speech can be classified into three different segments: (1) vocalic segment where formant structures are not only visible in the low-frequency region (say, below 3600 Hz.) but also dominate overall signal frequency distributions; (2) burst segment where signal energies are concentrated in the high-frequency region (say, above 4000 Hz.); and (3) mixed segment where the phonemes in both vocalic segment and burst segment are simultaneously present. Under this assumption, the notion of glottal-source spectra (gs-spectra) described in APT is more or less equivalent to the traditional definition of vocalic segment. The concepts of burst-friction spectra (bf-spectra) and mixed gs-spectra with bf-spectra are really not as difficult to perceive as they may seem, if we discuss them in simpler terms such as spectral peaks on the short-term spectral envelopes generated from standard LPC or FFT methods.

Let us look at both bf-spectra and gs-spectra mixed with bf-spectra through the examples shown in Figure 2.1 (a), and (b), respectively.



(a) Example plots of a set of *bf*-spectra.

(b) Example plots of a set of mixed *gs*- and *bf*-spectra.

Figure 2.1 Illurstration of sensory spectra of different types as shown in (a) a set of burst-friction (*bf*) spectra derived from the center of a /S/ segment and (b) a set of mixed spectra of glottal-source (*gs*) and *bf* dervied from the center of a /Z/ segment.

The running spectral envelope plots shown in Figure 5.1 are computed using a 24-pole LPC model from the center of /s/ segment in the word "SEVEN" and /z/ segment in the word "ZERO", each envelope representing the information contained in 512 data points. Now, we can describe this particular set of bf-spectra corresponding to /s/ segment by the facts that the spectra peaks of highest magnitudes across the entire frequency region are concentrated in a region around 8000 Hz. and there are no spectral peaks in the low-frequency region whose magnitude is "high enough" (note that we use very vague terms here to simplify our discussion). In the case of gs-spectra mixed with bf-spectra as shown in Figure 2.1 (b), there are three visible peaks in low-frequency region and the strongest peaks all occur in a very high frequency region (around 5000 Hz.).

Even if we have greatly simplified the concepts of auditory-sensory spectra which include the three different spectrums discussed here, we still run into problems in defining them in more precise terms, as in the case of using "high enough" to determine whether spectral peaks in low-frequency region are meaningful to our definition for bf-spectra. Obviously, we can not expect a fully automated recognition system to compute these spectra if we can not even define them clearly. In contrast, the gf-spectra is better defined and furthermore the major problem in gf-spectra generation is formant extraction in vocalic segment, which is more tractable although very difficult in actual practice.

Up to this point, we hope that the reader will agree that sensory path in the auditory-perceptual space described in APT can be specified

precisely for a vocalic segment of natural speech. However, the precise specification of formant extraction for vocalic segment does not necessarily guarantee that we can develop perfect formant extraction algorithms.

2.2.2 System Working Assumption

There are two main working assumptions under which the objectives of SWIS are established. First, it is assumed that information essential to phonetic perception contained in a vocalic segment is largely represented by the corresponding sensory path along with a few other features extracted from the segment. Second, we assume that it is possible to automatically generate a valid and sufficiently accurate sensory path for all possible vocalic segments. Here, by a valid sensory path we mean that a computer-generated sensory path matchs one generated manually by a human expert who knows the phonetic contexts of the segment. In other words, we have learned a great deal about formant structures of most vocalic segments from many years of basic research, and we can tell if a sensory path is valid or not once we have one and know its phonetic contexts.

The importance of the second assumption to this project cannot be overemphasized. After all, we are trying to develop some new technology based on APT to solve the problems in automatic speech recognition. For example, let's assume that most of the phonetic information contained in a vocalic segment can be near-exclusively represented by a path in a multidimensional space, but automatic generation of such a path from speech waveform is impossible. Research on the further analysis of paths in that space will not provide any solution to the old problems in

automatic speech recognition. This is simply because any feature-based ASR system will work only if these features can be automatically extracted from speech. Simply speaking, any brilliant idea is useless unless it is feasible. This is especially true in the history of ASR research.

Finally, this second assumption of this project, that is, the assumption that we can automatically generate a valid sensory path for any given vocalic segment with sufficiently high accuracy, may be hard to prove at this stage of our research. However, we have been quite successful in this effort in our past studies and other research efforts in this area, and, therefore, we are encouraged to start to investigate new problems arising from sensory-perceptual transformation and linguistic decoding processes, even while we continue to work on the problems associated with the generation of sensory paths.

2.2.3 SWIS Overview

The overall organization of SWIS shown in Figure 2.2 reflects our basic design philosophy discussed in this Chapter. The SWIS consists of two major components, namely, Sensory Path Generation (SPG) and Phonetic Decoder (PD). The SPG transforms input analog speech to sensory path. The SPG is considered as the front-end of SWIS although the digitization of analog speech and token editing are not part of SWIS. The details of SPG will be discussed in Chapter 4. Once sensory paths are generated they are passed to PD which in turn consists of two components, that is, Perceptual Path Generation (PGP) and Phonetic Parser (PP). The PGP will transform a sensory path into a perceptual path so that those phonetically unrelated attributes embedded in sensory path can be

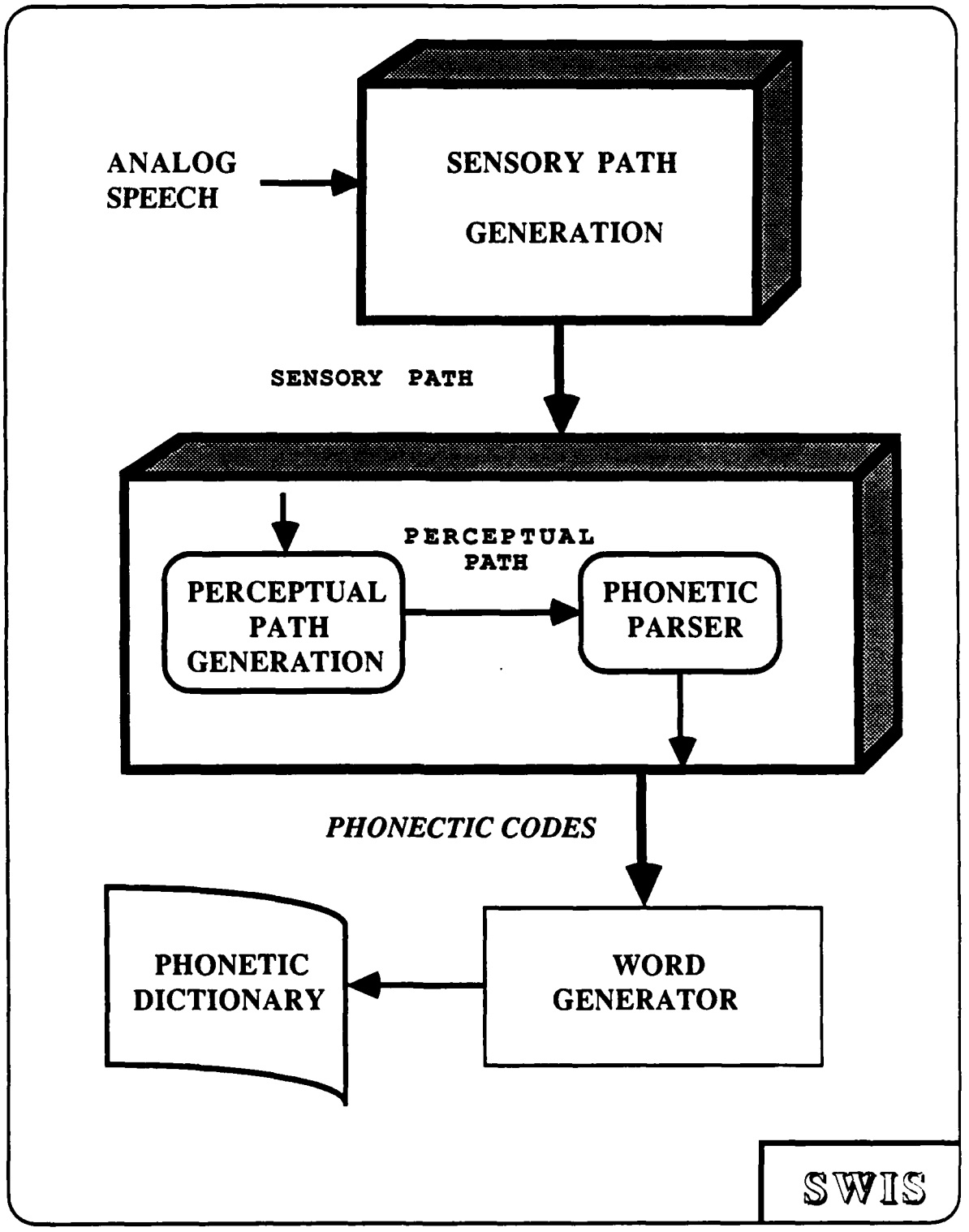


Figure 2.2 The major components of SWIS recognition system.

eliminated or at least be "turned" off. In essence, a perceptual path carries the same phonetic information as a sensory path does, plus those "highlighted" phonetically relevant attributes to facilitate subsequent processing in PD. The main function of PD is to parse a continuous perceptual path and generate a sequence of phonetic codes without explicitly locating phonetic segment boundaries. The details of PD will be described in Chapter 5.

Nearly all of the research efforts in SWIS project were to develop the SPG and PD software. In this project, the problem of lexical access, that is, determing the words spoken by the talkers from its acoustic-phonetic representation was given less attention. One reason was that the major emphasis was on the acoustic-phonetic transformation. The other was that the vocabulary DIGIT only includes eleven words. Lexical access was achieved by first establishing a variety of "phonetic" sequences associated with each word, that is, a phonetic dictionary. The word generator is simply a set of rules matching phonetic codes with the entries in the phonetic dictionary. Based on these rules, phonetic codes generated by SWIS from an utterance are matched against the entries in the phonetic dictionary. Thus, an utterance will be (1) correctly recognized if it is matched by one of the entries for that word, or (2) substituted or incorrectly recognized if it is matched by one of the entries for other words, or (3) rejected if no entry is found to match the codes. The detailed performance analysis on "recognition", "substitution", and "rejection" will be given in Chapter 6.

3. SIGNAL PROCESSING AND FEATURE REPRESENTATION

3.1 ACOUSTIC PREPROCESSOR AND LINEAR-PREDICTION ANALYSIS (LPA)

The digitized is procssed in the following way. (1) It passed through a highpass digitial filter of Chebychev designed with a cut-off frequency of 50 Hz. (2) Then a Hamming window of 25.6 milliseconds (ms), which is 512 samples at a 20 kHz sampling rate, is applied to the filtered waveform. (3) A high frequency preemphasis of the form, P(z)=1-(0.01*PR)z-1, where PR=0.98, is included in this process, and (4) Twenty-four autocorrelation coefficients are stored for each windowed waveform.

An analysis or feature frame has the same duration as the Hamming window. Frames are overlapped and spaced 3.2-ms (64 samples) apart. Therefore, a feature vector frame, centered within the 256-sample analysis buffer, is computed every 3.2-ms. A frame is the smallest entity in SWIS. From now on, we will frequently use "frame" in our discussion.

3.2 FEATURE REPRESENTATION

A frame consists of a set of features. Among these features are 24 LPA autocorrelation coefficients mentioned in the previous section. Table 3.1 gives a complete listing of features used in SWIS. Most of the features in Table 3.1 can be defined in more familiar terms in the literature. For those features obviously needing more explanation we provide a separate section for each.

We use the square-root of zero-th autocorrelation coefficient as total input signal amplitude (RO). Another familiar feature is zero-

TABLE 3.1 FEATURE LISTING

NAME	DESCRIPTION
A(0:23)	Autocorrelation Coefficient
R0	Total Signal Energy (Amplitude)
IZ	Zero-Crossing Rate
F0	Pitch or Fundamental Frequency
F1L	First Sensory Formant Low
F1LDB	Magnitude of First Sensory Formant Low
F1H	First Sensory Formant High
F1HDB	Magnitude of First Sensory Formant High
F2	Second Sensory Formant
F2DB	Magnitude of Second Sensory Formant
F3	Third Sensory Formant
F3DB	Magnitude of Third Sensory Formant
F4	Fourth Sensory Formant
F4DB	Magnitude of Fourth Sensory Formant
P1	Center Frequency of First Spectra Peak
P1DB	Magnitude of First Spectral Peak
PM	The Spectral Peak of Maximum Energy
PMDB	Magnitude of the Spectral Peak of Maximum Energy
SR	Sensory Reference
SF1L	Smoothed 1st Sensory Formant Low
SF1H	Smoothed 1st Sensory Formant High
SF2	Smoothed 2nd Sensory Formant
SF3	Smoothed 3rd Sensory Formant
X'	Perceptual Path X'-Coordinate in SLAB
Y'	Perceptual Path Y'-Coordinate in SLAB
Z'	Perceptual Path Z'-Coordinate in SLAB
V	Velocity of Perceptual Path (Log-Unit/Second)
A	Acceleration of Perceptual Path (Log-Unit/Second^2)
SI	Segmentation Index of Perceptual Path
ANGLE	Angle of Y'-Glide on Perceptual Path
DUR	Duration of Y'-Glide on Perceptual Path

Note: All of the formants are measured in Hz.
All the magnitudes are measured in decibels.

crossing rate (IZ) which is defined as the rate that samples change algebraic signs per second. F1L and F1H are low and high values of the first sensory formant, which in turn is defined as the first spectral prominence on the short-term spectral envelope. In non-nasalized speech, F1L and F1H are the same. In nasalized speech, there is usually only one spectral peak in the region where F1 and F2 normally reside. In this case, the lower edge of that spectral peak will be taken as F1L and the upper edge as F1H, i.e., splitting the first formant into two formants as proposed in APT. F1LDB and F1HDB are magnitudes of F1L and F1H in decibels. In actual implementation, in nasalized frames, F1L and F1H represent the low edge frequency location and the high edge frequency location of the first spectral prominence, respectively. For all non-nasalized frames F1L and F1H represent the center frequency of the first sensory formant. F2, F2DB, F3, F3DB, F4 and F4DB represent the center frequencies and the magnitudes of second, third, and fourth sensory formants, respectively. The extraction of sensory formants in both non-nasal frames and nasal frames is described in Chapter 5.

3.2.1 Fundamental Frequency (F0)

F0 is the pitch or fundamental frequency in Hz. F0 is generated by a cepstrally based, pitch extraction algorithm (82). The algorithm does not always generate a meaningful pitch value for each frame. For examples, sometimes it fails to locate two adjacent pitch pulses in a pitch buffer (32 msec.) and simply returns a zero. In this case we define F0 as the average pitch, which is computed from all the non-zero entries. If all the entries generated by the algorithm are zeros, then an estimated average F0 value is entered manually according to talker's

sex, to replace all the zero entries. A feature based time domain pitch tracker developed recently at CMU (67) claims a substantially better performance. We attempted to obtain a working version of this algorithm from CMU. However, the software is available for the PDP-11 and not readily adapted for the VAX systems used in our study. Pitch contours generated by the ILS system are very flat and provide no prosodic cues, and future works may be enhanced by better pitch extraction techniques.

3.2.2 Sensory Reference (SR)

As defined in APT, sensory reference (SR) is computed by the formula $SR=168(GMFO(i)/168)^{1/3}$, where GMFO(i) is the geometric mean of the talker's fundamental frequency at ith frame over the period starting from the first frame in an utterance up to the ith frame. This GMFO usually converges very quickly to the talker's average pitch after the first 10 to 15 frames in a vocalic segment. For simplicity, GMFO is replaced by F0 in SWIS because the ILS-generated pitch values are close to the talker's average pitch in most frames. Thus, SR is simply a function of F0 and varies slowly in time. It is mainly used together with F1L, F1H, F2 and F3 to define the auditory-perceptual space (APS) to minimize talker difference. See Section 2.1 for the detailed discussion on APS.

3.2.3 First Spectral Peak

First, we define the term "mean sensory reference" (MSR) in the equation 3.1 where AVGPITCH is the average of F0 contour.

$$MSR = 168(AVGPITCH/168)^{1/3} \qquad (3.1)$$

There are usually 12 or fewer spectral peaks on a spectral envelope derived from 24th order LPA model. P1 is simply the center frequency of the first peak on such an envelope. Note that P1 is not F1L or F1H although it often corresponds to the first sensory formant in non-nasalized frames. P1DB is the magnitude of P1 in decibels. If P1 computed in a particular frame is less than MSR, then P1 is reassigned to MSR while P1DB is not changed. Computation for P1 and P1DB is straight-forward. The first spectral peak can be easily detected by following slope changes on the spectral envelope. In case there are no peaks on the spectral envelope at a particular frame, i.e., the envelope is either a continuously falling or increasing curve throughout the entire frequency region, then P1 at that frame is simply assumed to be the value of the previous frame. If this happens in the first frame, P1 is assigned to MSR which is always defined. Figure 3.1 shows a set of the magnitude spectrums plotted in dB vs. frequency with the corresponding first spectral peaks marked with their center frequencies.

3.2.4 Maximum Energy Peak

Similarly to P1, the maximum energy peak is defined as the peak with the highest dB value on the spectral envelope. PM is the center frequency of such a peak and PMDB is its magnitude in decibels. Note that PMDB represents the maximum energy value in the entire frequency region (0-10 KHz.). The notion of maximum energy peak is developed based on the concept of "sensory-spectra" stated in the theory; they are defined as the first three "significant prominences" in the short-term spectral envelope of a speech waveform. In this sense, PM represents

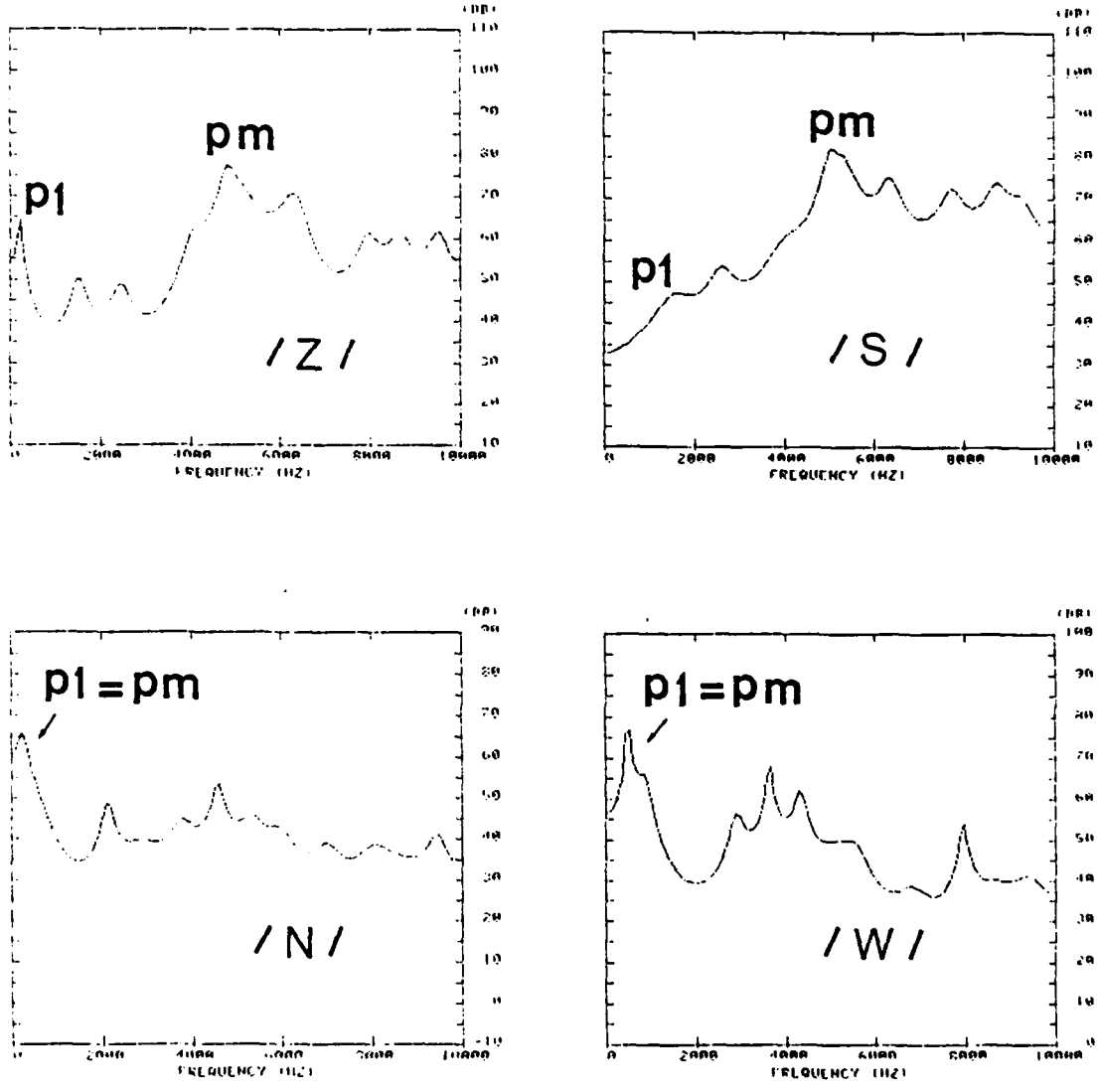


Figure 3.1 Illustration of selecting P1 and PM in different phonetic contexts.

the most "significant prominence" in the spectral envelope. Quite often, PM is equal to P1, especially in those frames where energy of speech signals concentrates in a low-frequency region, as shown in Figure 3.1.

To compute PM, SWIS first locates all the spectral peaks on the spectral envelope and then computes the magnitudes of each frequency component at its peak location. The highest magnitude calculated in such a way is taken as PM. For /Z/ spectrum shown in Figure 3.1, the peak of the highest magnitude (about 77 dB) on the envelope is located around 4300 Hz. In such a frame, PM will be equal to 4300 and PMDB will be equal to 77. Again, PM and PMDB will assume the values in the previous frame if there are no peaks found in a current frame. If this happens to the first frame, we simply use the maximum magnitude point on a spectral envelope to define PM and PMDB.

3.2.5 Smoothed Sensory Formants

Raw sensory formants F1L, F1H, F2, F3 and F4, (see Chapter 5.) are smoothed by a set of smoothing algorithms developed by the author. The details of the algorithms are also discussed in Chapter 5. SF1L, SF1H, SF2 and SF3 are the outputs of these algorithms when applied to F1L, F1H, F2, F3, and F4, respectively. Note that there are only 4 smoothed sensory formants because F4 is simply used in the smoothing algorithms to adjust SF3.

3.2.6 Perceptual Path

A perceptual path consists of a sequence of frames. Perceptual path is defined in SLAB coordinates (X', Y', and Z') rather than APS coordinates (X, Y, and Z'), see Appendix 9.1 for the details. Thus, the

location of the perceptual path in SLAB at any given frame can be expressed as P(X',Y',Z') where P is called the perceptual pointer and all the three coordinates are calculated at that frame in log-units. The perceptual path is generated through a sensory-perceptual transformation system which accepts sensory path as input. The details of the sensory-perceptual transformation will be discussed in Chapter 5.

3.2.7 Velocity and Acceleration

As defined in the previous section, the values of the perceptual coordinates in SLAB are related to the distance traveled along a perceptual path. In general, we are not only interested in the speed but also in the direction under which a perceptual pointer travels from frame to frame. The traditional term "velocity" is quite suitable to this context. But in the actual implementation here, only the magnitude is calculated as the absolute distance between two adjacent frames used, and the directional attributes of a perceptual path are ignored (i.e. up-down patterns along each dimension are not recorded at all). So, the velocity used in SWIS really represents the magnitude of velocity the perceptual path travels in SLAB space, which is usually called speed. APT uses "velocity" to describe how fast or how slow a perceptual path travels in APS. For consistency, we still use the term "velocity" rather than "speed" in SWIS so that there will be no confusion. However, the reader should keep in mind the actual formula used in velocity calculation. Velocity is measured in log-units per second while Acceleration, the time derivative of velocity, is measured in log-units per second squared. Clearly, both velocity and acceleration are variables in the time domain. So, velocity at ith frame is the absolute

distance between the i-1th frame and ith frame divided by step size (0.0032 second). The velocity at the first frame of any segment is defined as zero. The initial values of acceleration of the first two frames are handled in the same way.

3.2.8 Segmentation Index

First, we consider a perceptual path P(X',Y',Z') to be represented as three independent projections along X', Y' and Z' axes in SLAB space. We call each projection an X' curve, Y' curve, and Z' curve, respectively. Clearly, they are all the functions of time. Then, we simply locate all the peaks (i.e., local maximums) and valleys (local minimums) on the three curves. At each peak and valley, the maximum variation within time interval centered on that peak or valley are measured in log-units, i.e., the difference between the local minimum and the local maximum. The duration of such a region is set to a constant value (80 msec. involving 25 frames) for all the three curves. Thus, we have a pair, composed of a local minimum and a local maximum for each function at any peak or valley location. Finally, the three differences are added up with a weighting function to provide the values of the segmentation index (SI) at that peak or valley. For those non-peak or non-valley frames, their SI values are simply set to zero. The value of SI at ith frame is defined in equation 3.2,

$$SI(i) = w1*SIX(i) + w2*SIY(i) + w3*SIZ(i) \qquad (3.2)$$

where SIX(i), SIY(i), and SIZ(i) are the maximum variations of X', Y' and Z' coordinate values at the 80 msec. region centered at the ith frame. In actual design, w1 and w2 are set to one as a constant. W3 is

a function of Z' and varies inversely with Z' when ' ' is below 0.6. Thus, when a perceptual path travels through below the vowel-slab (such a Z'-dip usually indicates entering /r/ target or nasal targets), the corresponding frames will have high SI values. The formula used in calculation of W3 is given in equation 3.3., where Z(i) is the value of

$$w3(i) = \begin{cases} 1. & \text{if } Z(i) > 0.6 \\ 10.*(3.0/Z(i)) & \text{otherwise} \end{cases} \qquad (3.3)$$

Z'-coordinate at ith frame. Note that there is a break point for W3 at any frame where Z(i) is equal to 0.6. This discontinuity does not cause any problems because SI itself is not a continuous function, based on its definition. In fact, in most frames SI's have zero values because peaks or valleys on X', Y', and Z' are very limited.

3.2.9 Angle and Duration

Both angle (ANG) and duration (DUR) are the acoustic features used to describe a Y'-glide on a perceptual path. A Y'-glide is defined as a sequence of frames where their Y'-coordinate values constitute a non-decreasing sequence. Thus, if a Y'-glide starts at frame i and ends at frame j, then the following conditions must be met:

$$Y'(k) <= Y'(k+1) \qquad k=i, i+1, i+2, \ldots, j-1$$

The DUR of a Y'-glide is simply the total number of the frames in the Y'-glide, which includes the starting frame and the ending frame. Thus, if a Y'-glide starts from frame 20 and ends at frame 30 then its DUR is 11 (frames). Because the duration of each frame is fixed (3.2 msec) elapsed time of a Y'-glide can be easily calculated by multiplying DUR

by 3.2 msec. For example, if a Y'-glide has its DUR equal to 50, then its elapsed time will be 160 ms. long.

The ANG of a Y'-glide is a frequency-domain feature and needs more explanation. The ANG is calulated based on the locations of the starting point and the ending point of a Y'-glide. Figure 3.2 shows a Y'-glide starting from ith frame and ending at jth frame. Clearly, once we know the starting frame and the ending frame, we can use the X' and Y' coordinates associated with these two frames to compute ANG of that Y'-glide using standard arctan function.

Both ANG and DUR are associated with the starting frame of a Y'-glide because the ending frame number can be computed from the starting frame number and DUR. As such, the following assignments for ANG and DUR

ANG(23) = 75 (degrees)
DUR(23) = 50 (frames)

by Y'-glide detector indicate a Y'-glide has been found on a current perceptual path and the Y'-glide starts at frame 23 and ends at frame 72 (i.e., the duration is 160 msec long).

In order to minimize the computation time for ANG and DUR, the Y'-glide detector in SWIS first detects all the valleys (local minimum) on Y'-projection of a perceptual path and then works from there. Starting at each valley, Y'-glide detector follows Y' in time until Y' starts to fall. As a result, Y'-glide is a descriptor that can be used to easily locate the starting frame (valley frame) and the ending frame (falling frame).

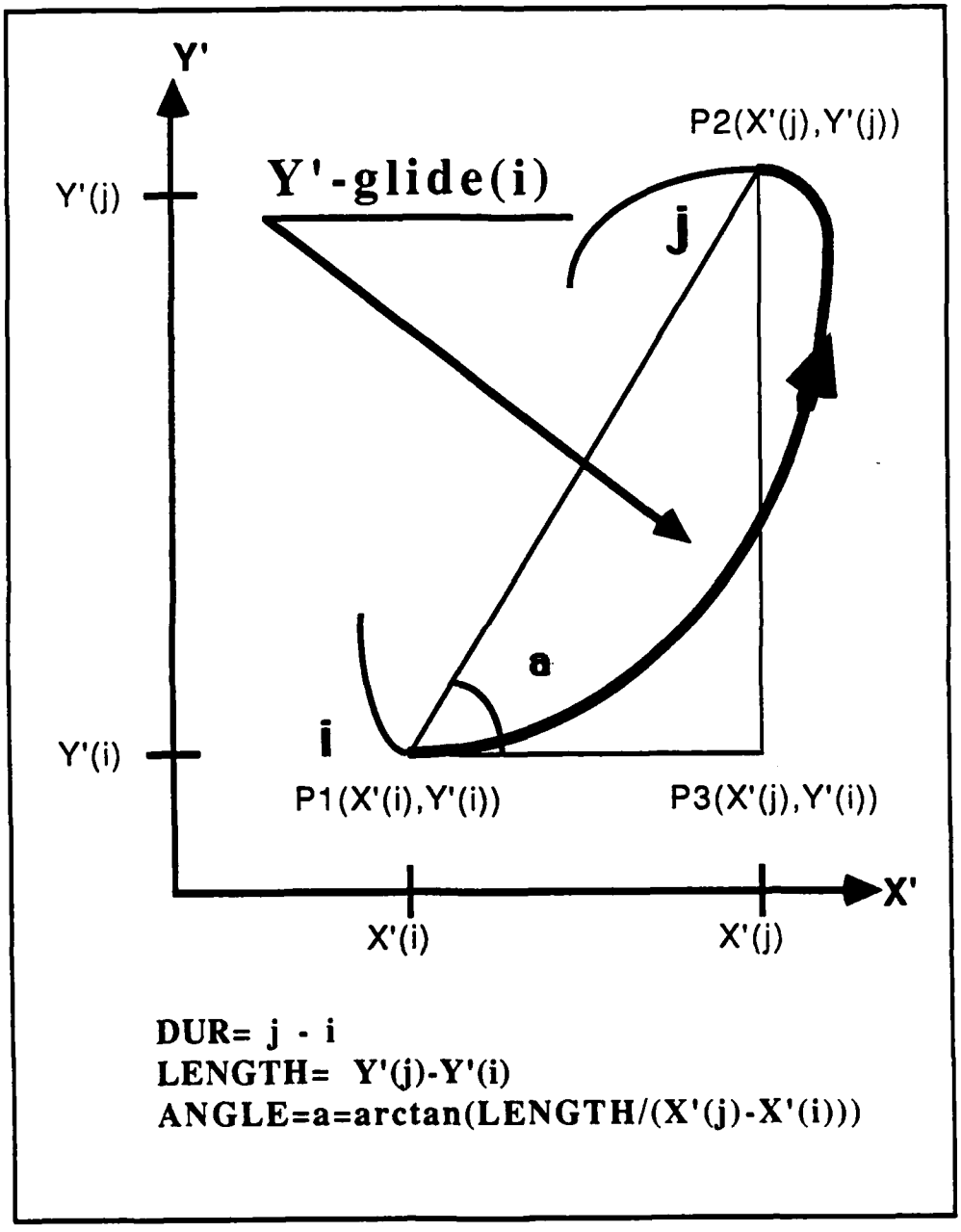


Figure 3.2 Illustration of a Y'-glide and its attributes. The Y'-glide starts at ith frame and ens at jth frame (darken segment).

During the initial testing of the Y'-glide detection algorithm on the train database (TRAIN), we noticed that there are some almost unperceptable downward fluctuations, which we call "trivial falls". They appear on Y'-projection of a perceptual path and are due to the lack of smoothness of the path. Those "trivial falls" often break the would-be non-decreasing sequence associated with a single Y'-glide. To overcome this problem, we developed a robust "fall" detection technique to make sure that Y' really starts to fall. So, a "fall" is identified only if there are three consecutive frames, i-1, i, and i+1 at the break point where Y'(i-1) is greater than Y'(i) which in turn is greater than Y'(i+1).

Because the concept of Y'-glide is introduced in SWIS only to capture the acoustic characteristics of a diphthong, several constraints are built into the detector. First, we assume that each diphthong has a minimum duration of 35 msec so that all the glides whose DUR are less than 10 (i.e. 32 msec) are simply ignored. That is, these DURs are reassigned to zeros. Second, Y'-glide should never overlap with a nasal segment and therefore it will be forced to end at any frame where Z'-coordinate value is less than 0.6. The details of nasal segmentation are described in the next Chapter.

Ideally, we wish to find a Y'-glide for each diphthong occurrence. Every Y'-glide would indicate that there is a potential diphthong segment at the region where the Y'-glide is found. Here, the key issue is to generate the minimum number of Y'-glides which cover all the diphthong occurrences. To avoid backtracking, SWIS looks for a diphthong only at each Y'-glide. This leads to one problem, that is, if

a diphthong does not appear as a Y'-glide then SWIS will miss it. However, almost all the diphthong occurrences do appear as Y'-glides when we test the algorithm with all the tokens in the database TEST1 and TEST2.

4. SWIS FRONT-END

4.1 INTRODUCTION

In SWIS, all the components including signal editing, digital filtering, LPC analysis, feature extraction, broad phonetic classification and sensory path generation are considered as part of "front-end" because they are relatively independent of vocabulary definition and use very little "high-level" knowledge. Our objective is to develop a front-end system which can be generalized to work with any vocabulary. That is, this front-end system should always locate the most stressed syllable of a word and then generate a sensory path from the vocalic segment containing the syllable.

The overall organization of the SWIS front-end is shown in Figure 4.1. Generally speaking, the SWIS front-end can be considered as a three stage process. In stage 1, acoustic features are extracted from digitized speech. Most of these acoustic features have been described in Chapter 3. Several features related to sensory path generation such as LMS, DPS, and SDGD (seen Figure 4.1) will be described in Section 4.3. Not every feature listed in Table 3.1 is utilized in the SWIS front-end. For example, features ANG and DUR for a Y'-glide are only used by diphthong detector - part of PD (Phonetic Decoder). In stage 2, an utterance is divided into several broad phonetic segments based on a set of features; this step is known as classification. The detailed phonetic classification techniques used in SWIS are presented in Section 4.2. In stage 3, a sensory path is generated for SGS (Strongest Glottal Source) segment identified by SWIS classification procedures. The

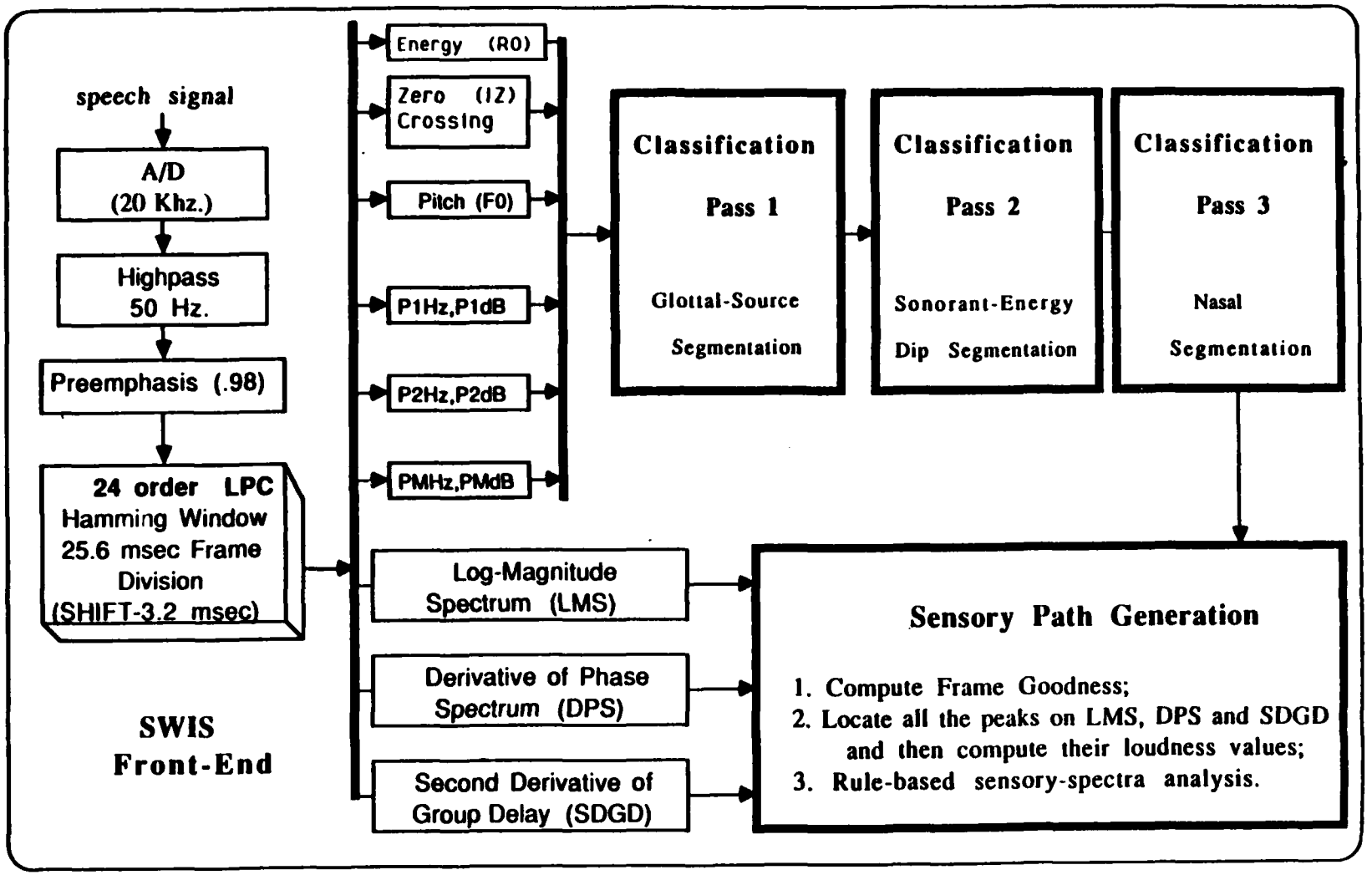


Figure 4.1 The block diagram of SWIS front-end system.

sensory path generation is the most important system component in SWIS front-end because it directly affects the overall system performance. Moreover, it is very difficult, if not impossible, to recover from many errors introduced in the sensory path generation once a sensory path has been generated. The detailed algorithms and feature parameters will be given in Section 4.3

SWIS does not generate any sensory path for non-SGS segments. Only a few simple attributes are created at each non-SGS segment, such as its duration, average zero-crossing rate (IZ) and total signal amplitude (RO). For each utterance, SWIS front-end creates a number of segments in different forms and passes them to PD (Phonetic Decoder) for further analysis. The SGS segment (there is one and only one SGS segment in an utterance) is represented by a sensory path while all other non-SGS segments are represented by a feature vector. However, the sensory path generator developed in this project could be upgraded to handle unstressed glottal source segments as well.

4.2 CLASSIFICATION

The concept of classification has been introduced in recent years to overcome the problems in detailed phonetic segmentation. It is very difficult, if not impossible, to directly divide the continuous speech signal into segments that correspond to detailed phonetic events. It is generally conceived that errors in detailed segmentation are inevitable based on our incomplete understanding of acoustic-phonetic correlates. Moreover, any error introduced in segmentation will directly affect the performance of subsequent processing and often propagate through an entire speech recognition system. Undoubtedly, any system that commits

itself to those inevitable errors at such an early stage is certainly not desirable.

The basic principle of classification is to make broad phonetic classifications at early stages and then to refine them later according to other system components' needs. Of course, the more classes there are, the more difficulty in classification. When the number of classes is equal to the number of phonemes in an underlying language, the classification is equivalent to the segmentation described earlier. In the past, several classification schemes have been proposed (20,43). For example, for all non-silent frames, speech can be classified in the following classes: A-like, I-like, U-like, Liquid, Nasal, Stop, and Fricative. For another example, an utterance can be defined as sequences of the following classes: vowel-like, obstruent, voiced-obstruent, silent, nasal, and sonorant-energy dip.

The main strategy used in classification in SWIS is to find a classification which can be extracted from the acoustic signal irrespective of local context, speaker characteristics, and other environmental variables. The number of classes initially should be kept small as long as the subsequent processing can proceed. As stated in APT, spectral analysis is defined in terms of glottal-source spectra with the special cases of nasalized spectra and burst-friction spectra. For simplicity, we will only use two classes initially to process sensory spectra, that is, glottal-source (GS) segment and non-GS segment. Another consideration in choosing a classification scheme is that the parameters used in classification must be reliably calculated from acoustic signals and their phonetic implications should be well

understood. Finally, the mistakes introduced in classification should be correctible. Based on these considerations, we shall use the following five classes to represent speech signals: (1) silent, (2) glottal-source, (3) sonorant-energy dip, (4) nasal and (5) unknown (wildcarding notation). The lexical representation of DIGIT defined in terms of these five phonetic classes is listed in Table 4.1. The last class is introduced to define those fuzzy boundaries between different phonetic events so that the mistakes made on other classes by subsequent spectral analysis procedures can be kept to a minimum. The classification was designed as a three-stage process. We will discuss the details of each stage in the next three sections.

4.2.1 Glottal-Source Detection: **Algorithm, G**

Waveforms generated from a glottal source show strong periodic characteristics when displayed in amplitude vs. time. The short-term spectral envelope of such waveforms often reveals the first three significant energy concentrations through their clearly-separated peaks. It is upon these three "prominent" spectral peaks that the sensory spectra are defined. In order to proceed with the sensory spectra analysis in the sensory path generation immediately following classification, we must locate all the glottal-source segments from an utterance.

The basic approach for detecting a glottal-source segment is first to label all the frames as silent or non-silent, then to determine which non-silent frame is qualified as a glottal-source frame. All those unqualified non-silent frames will be labeled unknown. Finally all the adjacent frames of the same type are merged into one segment.

Table 4.1 Lexical Representation of DIGIT in Broad Phonetic Classes

WORD	Lexical Entries	Length
ZERO	* GS *	3
ONE	* GS *	3
TWO	* GS *	3
THREE	* GS *	3
FOUR	* GS *	3
FIVE	* GS *	3
SIX	* GS * S *	5
SEVEN	* GS D GS N *	6
EIGHT	* GS * S *	5
NINE	* N GS N *	5
TEN	* GS N *	4

Total: 44

GS : Glottal Source Segement
N : Nasal Segment
D : Sonorant-energy Dip
S : Silent Segment
* : Unknown Segment

Before we introduce the detailed algorithm G for glottal-source detection, we need to state the problem in detecting the initial and final points of an utterance. Generally speaking, there is always a brief silence preceding and following utterances when they are recorded, so that no truncation occurs. Thus, detecting initial and final points is equivalent to detecting the end of the first and the begining of the last silent segment from the entire waveform. As shown in Figure 4.2 (a), a male utterance, "EIGHT," is initially divided by the algorithm into five segments [S][NS][S][NS][S], when S stands for silent and NS

stands for non-silent. For simplicity we will ignore all the initial and final silent segments when representing an utterance at such a broad phonetic level.

Generally speaking, R0 value is very large and IZ is very low at non-silent frames as shown in Figure 4.2 (a). In addition, this correlation is speaker-independent as far as our database is concerned. However, for those burst segments such as the final /t/ release in the word "EIGHT" (see Figure 4.2 (a)) their R0 values are very small and sometimes are indistinguishable from ones extracted from silent frames. Further, R0 values at those burst segments vary from talker to talker, even after normalizing R0 within the same utterance. This leads to the difficulty of the reliable detection of final unvoiced-stops of different talkers. In the case of word the "EIGHT", if the second NS (non-silent segment) is missed, then the intra-word pause, which is crucial to the identification of phonetic class orders associated with a word, is missed as well. On the other hand, if we could always recognize this intra-word pause, then we would have no problems in recognizing the word "EIGHT" because there is only one word in DIGIT satisfying the sequence of [NS][S][NS]. The point we try to make here is that more features are needed in classification even at this broadest phonetic classification level (only two categories, silent, non-silent).

The **algorithm G** employs multiple features to detect the transitions between different non-glottal source segments and the segments where glottal-source dominates their acoustic realizations. As shown in Figure 4.2 (b) by using R0 and IZ alone, it is very difficult to locate the boundary between fricative /Z/ and vowel /IH/. However, the sharp

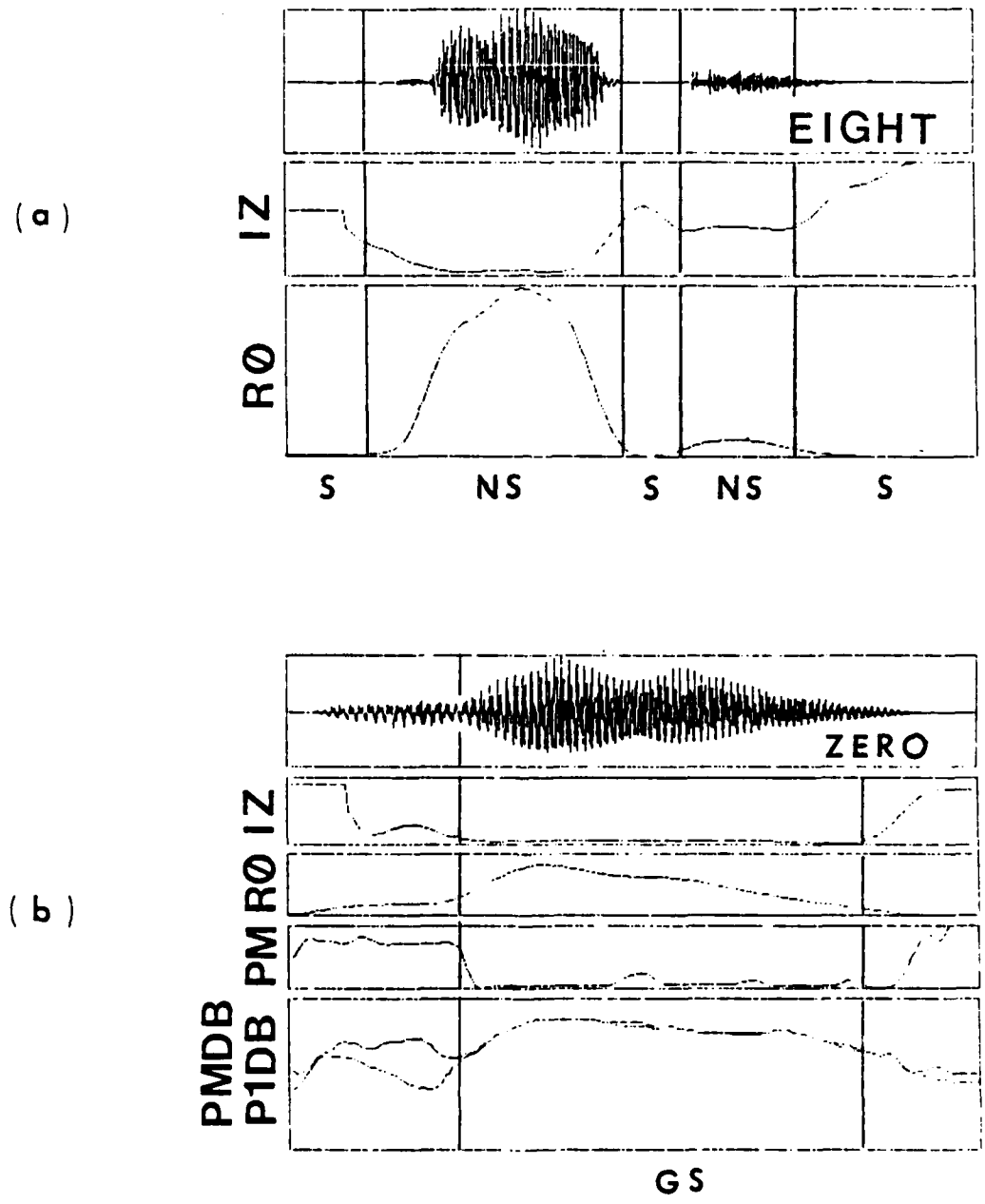


Figure 4.2 Illustration of multiple features based broad phonetic classification.

dip of PM and the changing difference between PIDB and PMDB from very large (30 db.) to zero clearly indicate the beginning of a glottal source segment. Now, we define the following notations used in G.

MAXR0:
the maximum value of R0 in the entire utterance.

MINR0:
the minimum value of R0 in the entire utterance.

SEGTYPE(I,1):
the type of ith frame

PEAKFN:
the peak frame number, i.e., R0(PEAKFN) = MAXR0

R0HIGH:
the high threshold for R0, defined as MAXR0 * 0.85

R0MID:
the middle threshold for R0, defined as MAXR0 * 0.55

VOWENG:
the minimum R0 value for vowel, defined as MAXR0 * 0.2

IZLOW:
the low zero-crossing threshold, defined as 5,000

IZHIGH:
the high zero-crossing threshold, defined as 6,500

The principles of the algorithm G are developed based on the following observations on the training database. That is, at a glottal-source frame, R0 should be relatively high (total signal amplitude is high due to voicing), IZ should be relatively low (periodic signals), formant structure must be visible at low-frequency region (below 3600 Hz.) and there should be a significant spectral prominence below 500 Hz. (pitch- related energy concentration). All the thresholds used in the algorithm were initially determined based on the analysis of the tokens

in the training database TRAIN (see Section 1.5.2) and later adjusted to fit the tokens in the first testing database TEST1 (see Section 1.5.2). The output of the algorithm G is a sequence of segments in type S (Silent), GS (Glottal Source) and * (unknown). As an example for the utterance "EIGHT" shown in Figure 4.2 (a), the algorithm generates the following segment sequence:

[*][GS][*][S][*]

The step-wise algorithm is the following:

Glottal-source Algorithm G

1. for every frame i, try to match the conditions from 2 through 7.
2. if R0(i) >= R0HIGH then SEGTYPE(i,1) = "GS"
3. if R0(i) >= R0MID and
   PMHZ(i) < 3,600 (Hz.) and
   PMDB(i) > 30 (db) then SEGTYPE(i,1) = "GS"
4. if R0(i) >= VOWENG and
   IZ(i) < IZLOW and
   P1DB(i) - PMDB(i) > - 15 (db) and
   PMHZ(i) < 4,200 (Hz) then SEGTYPE(i,1) = "GS"
5. if R0(i) < R0LOW and
   IZ(i) > IZ high then SEGTYPE(i,1) = "S"
6. if R0(i) < MINR0*3 then SEGTYPE(i,1) = "S"
7. if all above fail then SEGTYPE(i,1) = "*"
8. merge all the adjacent frames with the same type into one segment until all the frames have been processed. The number of segments are NSEG.
9. Locate the stressed segment S which contains the frame PEAKFN. Let V1 and V2 be the first and last frame of the segment S, respectively.
10. For i=1 To NSEG,
    SEG(i).AVGIZ = average IZ completed from segment i
    SEG(i).AVGR0 = average R0 completed from segment i

11. For i=2 To NSEG
    if (SEG(i-1).AVGIZ - SEG(i).AVGIZ)/10
      + ABS (SEG(i-1).AVGR0 - SEG(i).AVGR0) < 15
    then merge segment i into segment i-1;
    recompute average R0 and IZ for a segment i-1.

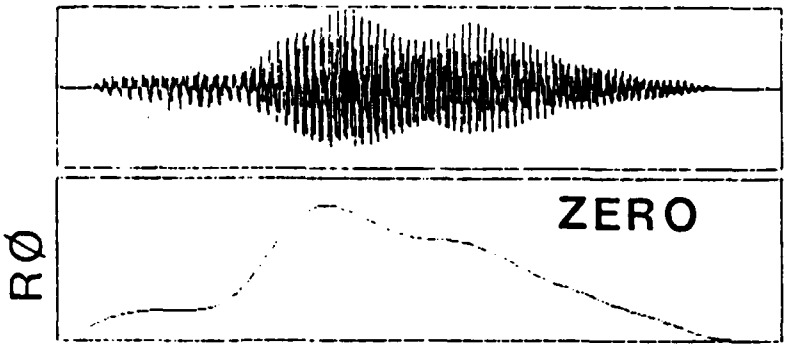
12. if there exists a pattern [*][GS][*] in the segment sequence generated so far and [GS] has 4 frames or less, then merge three segments into one [*].; recompute local averages of IZ and R0.

13. If there exists a pattern [*][S][*] in the segment sequence generated so far and [S] has 4 frames or less, then merge three segments into one [*]; recompute local averages of IZ and R0.

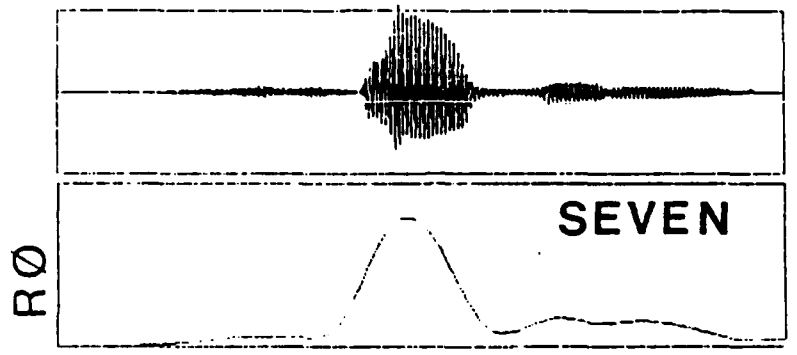
4.2.2 Sonorant-Energy Dip: **Algorithm SED**

If a glottal-source segment contains more than one syllable it is useful to find syllable boundaries so that the detailed phonetic segment labeling will be less difficult. One cue to reliably identifying multiple syllables in one glottal-source segment is known as sonorant-energy dip, which is a relatively-deep valley on smoothed signal energy contour (R0). This sonorant-energy dip is often caused by the CV syllable that follows the stressed syllable, especially when C is an /r/ or /v/. When we say this cue is reliable we really mean that it rarely generates false alarms if we define the threshold for a dip very conservatively.

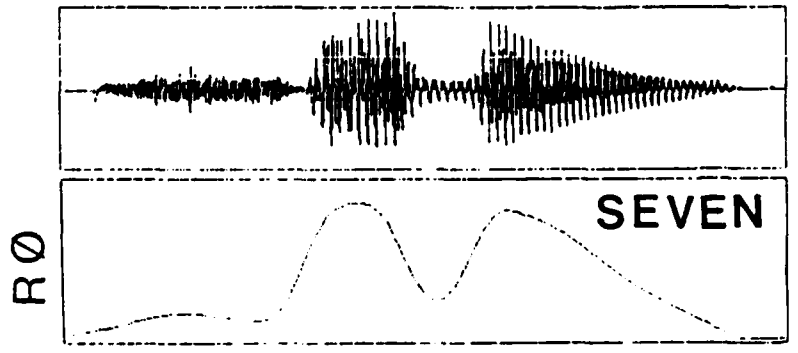
Any cue that is based on prosodic information such as sonorant-energy dip is sensitive to syllabic context and the difference in speakers' pronunciation. Figure 4.3 shows three sets of waveforms and their energy contour plots. As easily seen, only the energy contour on the bottom panel shows a visible dip at the beginning of the second syllable /V+AX+N/ in the word "SEVEN" (uttered by a male talker). In this particular example, R0 has two major peaks with a deep valley



(talker: MO, male)



(talker: LT, female)



(talker: MO, male)

Figure 4.3 Various appearances of sonorant-energy dip due to different syllabic contexts and talkers.

between them. We can calculate average peak/valley ratio based on the training tokens so that any valley falling between two peaks on R0 will be checked against the ratio threshold to determine if a sonorant energy dip has been encountered.

However, for some two-syllable words like ZERO their R0 contours do not always have such a dip at the boundary location between the two syllables. For example, the smoothed R0 plotted in the top panel in Figure 4.3 has no visible valleys although it is very easy to detect syllable boundary by examining the waveforms plotted above R0 contour. The smoothing procedure is responsible for the missing "dip". But if we did not use some kind of smoothing techniques on raw R0 contour, then we would have many small dips on R0.

A more complicated stepwise R0 smoothing technique has been developed by Huttenlocher and Zue at MIT (81) to preserve rise/fall patterns of R0 in the smoothing process. The reported performance on detecting sonorant-energy dip is very impressive. Unfortunately, the details of their algorithms used in the study were not presented in the report (because the algorithms are complex and not easily described except for the presentation of the Lisp code itself.)

Even with a better smoothing method, we still have the traditional problems associated with unstressed syllables. Usually, the signal intensity of an unstressed syllable is much less than one of a stressed syllable, which results in very weak peaks on R0 at unstressed syllable locations. As shown in the middle panel of Figure 4.3, the amplitude value associated with the second syllable is so weak that the corresponding frames are not even qualified as glottal-sources. That

is, the entire utterance is recognized as a two-syllable word with a pause between syllables.

Again, all the threshold values used in the algorithm SED are computed initially from the analysis of the tokens in TRAIN and later on adjusted to cover the exceptions introduced in the tokens in TEST1. Now we define the notations used in SED.

SNUM:
the stressed glottal-source segment number (computed by the algorithm G.)

NSEG:
total number of segments

SEG(i).FF:
the first frame of ith segment

SEG(i).LF:
the last frame of ith segment

SEG(i).TYPE:
the type of ith segment

SEG(i).*:
all the attributes associated with the ith segment

```
V2 = SEG(SNUM).LF   { the last frame of segment SNUM }
```

Algorithm SED

```
1.  Find all local valleys on the R0 Contour of S.
    VL(i) is the frame number of ith valley.
    { assume that NV valleys are found in this step. }

2.  for i=1 To NV execute 3 to 11

    3.  if 20log(MAXR0/R0(VL(i)) < 6 (db) then ignore this valley.
        go to 11
        { if the current valley is not deep enough, then
          it is not associated with a sonorant energy dip.  }

    4.  Set DL to the first frame of ith dip, initialized to VL(i)
```

```
5.  DO  DL=DL-1  until (R0(DL)-R0(VL(i)) > 3,500)
    { try to find the starting point of this valley }

6.  Set DR to the last frame of ith dip, initialized to VL(i)

7.  DO {
       DR=DR + 1
       if DR > V2 then ignore this dip, go to 11
       }
    until (R0(DR)-R0(VL(i)) > 3,000)
    { try to find the ending point of this valley }

8.  for i=NSEG To S Step -1
    SEG(i+2).* = SEG(i).*

9.  Break SNUM into three segments
    SEG(SNUM).LF = VL(i)-1
    SEG(SNUM+1).FF = DL
    SEG(SNUM+1).LF = DR
    SEG(SNUM+1).TYPE = dip
    SEG(SNUM+2).FF = DR+1
    SEG(SNUM+2).LF = V2
    SEG(SNUM+2).TYPE = glottal-source
    V2 = DL-1

    recompute average IZ and R0 for all the three segments

10. Goto 12
     { quit as soon as the first sonorant-energy dip is found  }

11. Complete processing ith valley

12. Terminate the execution of SED.
```

Note that SED will terminate when the first dip is found. On the other hand, SED may terminate with no dip found. Generally speaking, the algorithm SED will divide a SGS segment into three segments (two glottal-source segments plus a dip segment between them) if there is a sonorant-energy dip found in the SGS segment. Let's name these three segments in the order they appear in a SGS segment as left-GS, dip and right-GS, respectively. Clearly, there must be one GS segment

containing the stressed syllable in that utterance, either one to the left of the dip segment or one to the right. We assume that the GS segment containing the peak frame (where R0 has the maximum value in an entire utterance) will include the stressed syllable and is renamed as a current SGS segment. As such, the sensory path generator will only process this new SGS segment and will not generate any sensory path for the non-stressed glottal-source segment. For example, an utterance for the word "SEVEN" might be divided into three glottal-source segments, first for /EH/ in the stressed syllable /S+EH/, second for /v/ in the second syllable /V+DX+N/ and third for for /DX+N/. As defined above, only the sensory path for /EH/ is generated for detailed phonetic labeling. The other two segments will be labeled at a high-level, i.e., as a "DIP" segment and as an "Unstressed Glottal-Source". No further analysis is performed to find the identities of "DIP" and "Unstressed Glottal-Source" because the resulting phonetic code sequence listed below will match one and only

[EH][DIP][Unstressed Glottal-Source]

one word (only the word "SEVEN" might contain such a sequence) in DIGIT.

4.2.3 Nasal Segmentation: **Algorithm, N**

A SGS segment might contain one or more nasal segments. For example, the SGS segment for the word "NINE" contains a nasal-initial segment and a nasal-final segment. If we can locate a nasal-initial or a nasal-final segment or both from an underlying SGS segment, then we will reduce substantially the search space (for example, in the case of "NINE" only the words containing [Nasal][Vowel][Nasal] subsequence will be considered). Moreover, once a nasal segment is located, the

subsequent sensory path generation will be less difficult because now it affects non-nasalized vocalic segments only.

We assume that there is at most one initial nasal segment and at most one final nasal segment within a stressed glottal source segment. This assumption is reasonable because a SGS segment should contain at most two syllables, one of which is the stressed syllable in the word spoken. The goal of the **Algorithm N** is to find if there is an initial nasal segment and/or a final nasal segment in a stressed glottal source segment. Once a nasal segment is found, all the frames in the segment will be marked as nasal frames so that sensory formants in these nasal frames will be generated based on the rules that differ for non-nasal frames.

The basic approach to nasal segmentation in SWIS is to identify temporal and spectral characteristics associated with the transition into/from a nasal segment from/to an adjacent vowel. For example, by examining P1 contour, we notice that there is a sharp rise at nasal-offset location where the formant pattern for the following vowel becomes clearly identifiable. In the case of nasal-final, we see the opposite pattern — a sharp falling on P1 contour at nasal-onset location where the formant pattern for the vowel undergoes significant changes. In addition, following the sharp fall at nasal-onset location, there is a steady-state region on P1 contour where P1 normally stays below 300 Hz. and has very little change (flat curve) within the region.

Sometimes, it is very difficult to define "sharp-fall" or "sharp-rise" when talkers' pitch have low resonance. Unlike many other problems in speech recognition, this difficulty is greater for male than

for female utterance. In case of a female utterance, such a "sharp-fall" or "sharp-rise" often translates into a 400 Hz. drop or rise on the P1 contour within a region of 30 to 80 msec, and, there are no other transitions of this rapid rate except for nasal-initial or nasal-final in CV, CVC, and VC syllable. In other words, we can use this rising or falling location to determine when a nasal segment ends (entering a following vowel segment) and when a nasal starts (exiting from the preceding vowel segment). However, in the case of a male utterance, such a rise or fall is not very evident. This is especially true for the word "NINE" because P1 at the last 50 msec. of diphthong /AY/ is very low (about 300 Hz. for a male utterance) and P1 at the following nasal segment is about 200 Hz. So, there is little room for P1 to display its "drop" pattern. All we can see from P1 at the nasal-offset region is a gradual transition from about 300 Hz. to 200 Hz., which can be found in many words which end with a vowel such as "THREE" or "ZERO".

In order to compensate for the lack of unique temporal and spectral patterns associated with nasal-transition, we add an energy concentration test in frequence domain at the nasal nucleus. Of course, we have to locate a nasal segment before we can find a nasal nucleus. But if we know where there is a nasal segment, we don't need this nasal nucleus. The way we solve this conflict is to use a two-pass check in the Algorithm N. In pass 1, we tentatively mark a segment as nasal based on the temporal and spectral patterns associated with nasal transition. In pass 2, we apply the energy concentration test on a tentatively marked nasal segment to make a final decision.

The energy concentration test is mainly to check whether the magnitude of the first spectral peak is significantly higher than that of the second spectral peak at the frames in a nasal nucleus (i.e., the center of a nasal segment). This test eliminates many false-alarms of marking a normal vowel segment as a nasal segment because the magnitudes of low-frequency spectral peaks on non-nasal segments usually do not vary much. There is only one exception for the previous statement, that is, semivowel /w/ does have a significant difference between the magnitudes of its P1 and P2. We would expect that some /w/-segments will be mislabeled as nasal segments in the final testing.

Again, the nasal detection algorithm N is just an experimental method to solve the problems in the identification of nasal segments. By no means, are we trying to to develop a powerful nasal detection algorithm (it alone might require another Ph.D thesis to accomplish). All we are trying to demonstrate is that it can be done in this broad framework. Furthermore, once we identify a nasal segment, either a nasal-initial or nasal-final, we make no attempt to do detailed labeling, that is, to decide if this nasal segment is a /N/, or /M/, or /NX/. This is simply because there is no word containing /M/ or /NX/ in DIGIT. Even for a large vocabulary IWR, we believe that such a difficult process should be delayed. For example, if there are two or more word candidates represented in the same phonetic sequence [Nasal][AY][N], (they will match both word "NINE" and "MINE"), then, in order to make the final decision on whether that spoken word is NINE or MINE, the system must find out if the nasal segment is a /M/ or /N/. In

continuous speech recognition, we would use semantic constraints to add word recognition for the example discussed above.

We define the following notations used in **Algorithm N.**

IZHIGH: same as used in Algorithm G.

NBUF(i):
A nasal buffer centered at ith frame (7 frames).

NFMAX:
Maximum value of P1 allowed in a nasal frame.

P1:
First spectral peak contour (see 3.2.3)

P2DB(i):
Magnitude of the second spectral peak on the spectral envelope generated at ith frame (dB).

RONMAX:
Maximum RO in a nasal frame.

SNUM:
the stressed glottal source segment number. (see 4.1.1)

V1:
the first frame of SNUM.

V2:
the last frame of SNUM.

Note that all the statements enclosed in "{" and "}" are comments for those steps in the algorithm which need further explanation.

ALGORITHM N.

1. Mark all the frames in SNUM as non-nasal frames.

2. { initialize the variables used in the algorithm. }
   Set P1_drop to false; Set P1_fall to false;
   Set P1_steady to false; Set NINIT_found false;
   Set NFINAL_found to false;
   NI.FF = V1; NF.LF = V2;
   NFMAX=GMTSR*1.5; RONMAX=6,000

```
3.  For i = V1+3 To V2-4 execute 4 through 12

    4.  { In a nasal segment, IZ should be low and
          the center frequency of the first spectral
          peak is usually less than 500 Hz }

        If IZ(i+4) > IZHIGH or P1(i) > 500 goto 12


    5.  { compute the parameters used to determine
          if a steady-state is found on P1 contour }

        P1AVG = average P1 in NBUF(i)
        P1LOW = minimum of P1 in NBUF(i)
        P1HIGH = maximum of P1 in NBUF(i)
        P1DIF = P1 at the first frame in NBUF(i)
                - P1 at the last frame in NBUF(i)


    6.  { mark the steady-state region on the P1 contour }

        If P1HIGH - P1AVG < 20 and P1AVG - P1LOW > 20
        Then
             P1_steady is true;  P1S = I
        Else
             If I=P1S > 10 then P1_steady is false.
        End if


    7.  { locate the end of a nasal-initial segment }

        If P1_steady is true and
           P1 is continuously rising in NBUF(i) and
           P1HIGH - P1LOW > 75 and
           NINIT_found is false and
           P1(i) < 300
        Then
           NI.LF = i + 3; NINIT_Found is true.
        End if


    8.  { mark the sudden drop point on the P1 contour  }

        If P1DIF > 250 Then
             P1_drop is true; P1D = i
        Else
             If I - P1D > 20 then P1_drop is false
        End if
```

```
9.  If P1DIF > 50 and P1 is continuously falling
    Then
        P1_fall is true; P1_fall_at = i
    End if

10. If I - P1_fall_at > 18 Then P1_fall is false.

11. {  A steady-state region following a sudden
       drop on P1, then it indicates the begining
       of a nasal-final segment.  }

    If (P1_fall is true or P1_drop is true) and
       (P1(i+3) < NFMAX or P1_drop is true) and
       RO(i+3) < RONMAX and
       P1_steady is true and
       V2 - i > 10
    Then
       NF.FF = i+3;  NFINAL_found is true.
    End if

12. If NFINAL_found is true goto 13

13. If NINIT_found is true
    Then
       for i = NI.FF, NI.LF
            mark ith frame as a nasal frame
       End for
    End if

14. {  varify if it is really a nasal-final segment. }

    If NFINAL_found is true
    Then
       { find the center of this nasal-final segment  }
       CFN = min(NF.FF+8, V2-2)

        D12 = average( P1DB(i)-P2DB(i) )
              where i is in the region of [CFN-2,CFN+2];

        If D12 > 10 dB then
            For i=NF.FF,NF.LF
               mark ith frame as a nasal frame;
            End for
       End if
    End if
```

4.2.4 Discussion

Tables 4.2 and 4.3 summarize classification errors on the initial test using the tokens in the TRAIN database and the TEST1 database, respectively. Only those tokens whose class sequences do not match predefined template sequences (see Table 4.1) are listed. In the current implementation of SWIS, nasal detection **Algorithm N** applies only to a stressed glottal source segment. Therefore, no nasal segment is found in an unstressed glottal source segment such as the second syllable /V + AX + N/ in "SEVEN".

The two types of classification errors, namely, Inserted/Missed and Mislabeled, are defined to determine the performance of SWIS classification procedures. In the former, the resulting segment sequence has either more numbers of segments or less numbers of segments compared with the predefined class templates for each word in DIGIT. For example, an extra nasal segment is inserted after glottal-source segment for the utterance "THREE" by talker JH (see Table 4.3). In the latter, the number of segments of a word generated by SWIS classification programs is the same as defined for that word but there are one or more segments having different labels than the class templates defined for DIGIT. For example, the third segment for the first utterance for talker TS is labeled as [*] (unknown segment) rather than [D] (sonorant-energy dip segment) because the algorithm D fails to identify that the segment is a sonorant-energy dip. Obviously, such a mistake is not fatal because the algorithm D does recognize the utterance is a two-syllable word and the sequence [*][GS][*][GS][*] can only be matched by the word "SEVEN".

Table 4.2 Classification Errors Computed from the Training Database TRAIN

WORD	ID	SWIS GENERATED SEQUENCE	I/D*	ML**
THREE	LT-2	* GS N *	1	
SEVEN	LT-1	* GS * GS *	1	1
	LT-2	* GS * GS *	1	1
total			3	2

Error rate (%) = (5/(44*2*2))*100 = 2.7%

* I/D: Number of Inserted/Deleted Classes
** ML : Number of Mislabled Classes

Table 4.3 Classification Errors Computed from the Training Database TEST1

WORD	ID	SWIS GENERATED SEQUENCE	I/D*	ML**
ZERO	JH-1	* GS N *	1	
ONE	TS-1	* GS * GS N *	1	
THREE	JH-2	* GS * N *	1	
SEVEN	JH-1	* GS D GS *	1	
	JH-2	* GS D GS *	1	
	TS-1	* GS * GS *	1	1
	TS-2	* GS * N *	1	1
EIGHT	TS-1	* GS *	2	
NINE	JH-1	* GS * N *	1	
	TS-1	* GS *	2	
	TS-2	* GS *	2	
TEN	TS-1	* GS *	1	
	TS-2	* GS *	1	
total			16	2

Error Rate (%) = (16+2)/(44*2*2)*100 = 10%

As seen in Tables 4.2 and 4.3 most errors are due to the fact that **Algorithm N** fails to identify an initial and final nasal segment. However, all the missed nasal-final segments can be easily recovered after key phonemes in the stressed glottal-source segment are identified, so that overall recognition accuracy will not be affected.

Because the overall error rate for broad phonetic classification procedures from our initial design is less than 10%, we felt at that time that we should continue the software development for sensory path generation rather than spend more time to modify the algorithms used in classification. Therefore, only minor changes were made on the algorithms **G**, **D** and **SED**, based on the problems discovered from running SWIS with the tokens in TEST1. After this modification, all the errors made on TRAIN and TEST1 were corrected, except for "intentionally missed" nasal segments in unstressed glottal-source segments (because SWIS does not look for a nasal segment within an unstressed syllable segment).

Appendix 10.2 provides a complete set of classification plots of DIGIT. For each word in DIGIT, two plots showing the final classification results are given, one for each talker in the training group. All the utterances used for plots are from the train database TRAIN.

4.3 SENSORY PATH GENERATION

4.3.1 Introduction

The central problem in sensory path generation is formant extraction. Before we start to review the problems in formant extraction, which are well-known by the speech recognition community, we

would like to make clear a distinction between two different concepts, namely, sensory spectra and formants. As commonly defined, formants are the resonant frequencies of the vocal tract. Therefore, the fundamental frequency of the glottis is defined as F0 (pitch), and the first, second and third resonant frequencies of the vocal tract are defined as F1, F2 and F3, respectively.

Note that formants are the properties associated with the production of acoustic signals. In other words, these resonant frequencies of the vocal tract are always there regardless of whether listeners perceive them or not. On the other hand, sensory spectra are defined in terms of their strongest prominences that are important to listeners' perception for the underlying sounds. Therefore, sensory formants SF1, SF2, and SF3 are often not equal to F1, F2, and F3, respectively. For example, if the second resonance of the vocal tract at a given time is very weak relative to the third resonance, then SF2 will take the center frequencies of F3 rather than F2.

In SWIS, the sensory path generation is actually a two-stage process. In stage 1, the traditional formant extraction is performed to obtain F1, F2, F3 and F4. In stage 2, the resulting formants are weighed to decide the assignments of sensory formants. We will review the various past methods of formant extraction studies in the past in section 4.3.2. In section 4.3.3 we will present three types of spectra used in our analysis of sensory spectra. In section 4.3.4 the algorithm for sensory path generation (SPG) is described. In section 4.3.5 we will discuss the performance of the SPG algorithm based on the test conducted with the database TRAIN described in section 1.5.1.

4.3.2 Review of Formant Extraction Methods

Several methods have been proposed to automatically estimate the center frequencies of vocal tract resonances, that is, formant values. The resonant nature of a speech signal is best manifested in the short-time spectrum in which several spectral peaks can be found. Except for a few earlier attempts, most of the methods published in the literature are based on the short-time spectrum in various representations. Examples of these earlier efforts are the spectrum-sampling formant extractor developed by James Flanagan in 1955 (83), in which he used 36 analog bandpass filters, and the pitch-synchronous time-domain estimation of formant frequency and bandwidth, proposed by Elliot Pinson in 1963 (84).

From the latter 1960s to today, the research in formant extraction has been mainly concentrated on the development of peak-searching based algorithms with different spectral representations. In 1969, Schafer and Rabiner (85) reported a peak-searching algorithm on smoothed spectra obtained by using the Cepstral method. This method has problems because it recognizes spurious peaks and fails to distinguish merged peaks. Linear-Prediction Analysis (LPA) of speech has made formant estimation more tractable since spurious peaks are rare (69). However, a comparison study (86) of LPA formant-extraction algorithms indicates that the merged peaks on LPA spectral envelopes can not be easily resolved.

An obvious method for extracting formants from LPA spectra would be to solve the roots of the polynomial whose coefficients are given by the LPA coefficients, but this is computationally complex and requires high

precision complex-number arithmetic. Similarly, an analysis-by-synthesis method for solving a set of non-linear simultaneous equations using a least-square fit was proposed to obtain formants directly from FFT spectra (87). Again, this method is computationally expensive. In addition, there are the problems arising from the fact that the iterations involved in the process are not always convergent. Further, the algorithm requires the fixed bandwidth for each formant to be extracted and, therefore, may lead to potential problems when used in speaker-independent speech recognition systems.

In recent years, several methods have been proposed to solve the problem of merged peaks. One of them (86) is to use the Second Derivative of the Log Magnitude Spectrum (SDLMS) computed from the LPA method. Yegnanarayana (88) has proposed another method of identifying closely spaced formants. The method uses the Derivative of the Phase Spectrum (DPS) instead of the magnitude spectrum. The latest technique used for formant extraction was reported by Reddy (71). It basically combines the three successful formant extraction algorithms developed in the past to extract formant with very high resolution. Almost all the algorithms reported in the literature have been tested with synthetic speech. An exception is the McCandless' algorithm (69), where 50 sentences of natural speech were used in the final testing, but the number of talkers involved in this test was not stated. In addition, none of the reports mentioned above give the details of their implementations, for example, such as listings of their computer programs. Based on our own evaluation of the known algorithms in terms of performance, complexity and efficiency, Reddy's algorithm was

considered to be the best at the time we started to implement the program for sensory path generationin SWIS. With a few modifications, the algorithm used in SWIS formant extraction is largely based on Reddy's methods.

4.3.3 Spectrum Representation

Three forms of spectra are used in SWIS to represent spectral information, namely, Log-Magnitude Spectrum (LMS), Derivative of Phase Spectrum (DPS), and Second Derivative of Group Delay (SDGD). All the spectra are derived from 24 autocorrelation coefficients computed from preemphasized speech waveforms. The choice of 24 predictor coefficients was reached through experimentation. It was found, with fewer coefficients, the merging formants of certain sounds, such as /AX/ and /ER/, were presented by only one peak at the region where F2 and F3 usually reside. With 24 coefficients almost all the first three formants of pure vowels appear as individual peaks at their steady states. But spurious peaks are common, especially at the transition region between vowel nucleus and adjacent consonants. However, spurious peaks may be identified at a later stage of formant analysis whereas a missing peak is an irrecoverable problem.

Once the 24 coefficients are available, the LMS can be computed from these coefficients using an N-point FFT pruning algorithm. Generally speaking, N can be arbitrarily large to increase frequency resolution at the expense of computation time. Traditionally, N has been limited to a small number, say, 128 or 256 at most, since the main concerns are that high-order FFT is very expensive to implement in real time. Because of the advanced research in special digital signal

processing (DSP) chips, we no longer need to be concerned with the cost of DFT real-time implementation. For example, 1024-point FFT can be computed in less than 1 msec by some DSP chips, which is faster than real time for speech recognition applications. For this reason, we have chosen N=1024, resulting in approximately 20 Hz spectral resolution. This means that, (1) formant values are accurate to within 20 Hz and (2) any two spectral peaks which are more than 40 Hz apart can be resolved. The accuracy is sufficient for our needs and we find that it is very rare to have two spectral peaks separated by less than 50 Hz. In addition, the DFT routine used in SWIS is implemented with a table lookup method, so that 1024-point FFT does not slow down the simulation process.

We now present the detailed steps of computation for each spectrum mentioned above.

Log Magnitude Spectrum (LMS)

(1) Compute 1024-point complex DFT of $\{1, a_1, a_2, \ldots, a_{24}, 0, \ldots, 0\}$ where a_i is the ith autocorrelation coefficient. The output of DFT routine is a sequence of complex DFT coefficients

$$(a_1, b_1), (a_2, b_2), \ldots, (a_{1024}, b_{1024})$$

(2) Compute LMS at each spectral component location as follows:

$$LMS(i) = -8.686\ln((a_i^2+b_i^2)^{1/2}) \quad (4.1)$$

where i=1,2, ..., 512 and the unit of LMS is in dB and -8.686 is a coefficient used in the conversion from natural log to common log.

Derivative of Phase Spectrum (DPS)

(1) Compute the phase spectrum from complex DFT coefficients using the standard arctan function subroutine as follows:

$$PS(i) = \arctan(b_i/a_i) \quad (4.2)$$

where i=1,2, ..., 512.

(2) The derivative of the phase spectrum (DPS) is then obtained by computing the difference between the successive points in the frequency domain.

$$DPS(i) = PS(i) - PS(i-1) \quad (4.3)$$

where i=2,3, ..., 512, DPS(1) = 0 and the unit of DPS is in angle-frequency.

Second Derivative of Group Delay (SDGD)

(1) Compute SDGD from DPS as follows:

$$SDGD(i) = -DPS(i-1)+2DPS(i)-DPS(i+1) \quad (4.4)$$

where i=1,2, ..., 512 and SDGD(1) = 0.

4.3.4 Identification of Sensory Formants

In this section we present the algorithm used in SPG to locate sensory formants: SF1L, SF1H, SF2 and SF3. As we mentioned earlier, the algorithm is based mainly on Reddy's method. In Reddy's algorithm, all the formants are derived from the local maximums in SDGD using a complicated pole-verification method to achieve high resolution of the resulting formants. However, in SPG we need to frequently renumber spectral peaks to approriately map these spectral peaks to sensory formants. Therefore, we need a very efficient method of performing initial formant estimation based on the short-term spectra. Bearing this in mind, we decided to develop an algorithm based on the peak analysis on all the three spectra. That is, three sets of spectral peaks, one for each spectra type - LMS, DPS, or SDGD), are all taken into consideration to determine the locations of sensory formants at each analysis frame.

Several terms used in the following algorithm need special explanation. One of them is MDIST, which is based on the concept of sensory spectral bands proposed in APT (see Appendix 10.1). MDIST is defined as the log-ratio of the frequency location of a spectral peak to MSR (see Section 3.2.3). For example, if MDIST(1) is such a ratio for the first spectral peak located at 600 Hz, and MSR is calculated as 150 Hz then MDIST(1) will be computed as follows:

$$\text{MDIST}(1) = \log(600/150) = \log 4 = 0.602$$

The use of MDIST in the SPG algorithm will eliminate any obvious errors in the peak assignment; for example, a peak located at 2000 Hz

should not be assigned as SF1 because the first sensory formant is generally below 1800 Hz in most phonetic contexts. The ranges of MDIST were originally proposed in APT as (0.,0.8) for SF1, (0.6,1.2) for SF2 and (1.0,1.4) for SF3. Based on our testing data they have been modified as shown in the following equations.

$$0.0 < MDIST(1) < 0.8 \tag{4.5}$$

$$0.56 < MDIST(2) < 1.2 \tag{4.6}$$

$$0.86 < MDIST(3) < 1.4 \tag{4.7}$$

Another special term used in the SPG algorithm is frame goodness (FG). Generally speaking, if a frame is located in a region where signal amplitude is high, the phonetic information contained in that frame is considered more reliable than that in an region where signal amplitude is low. To obtain a uniform definition of "high signal amplitude," we set a reference amplitude as the maximum value of R0 in a vocalic segment. Then, if the R0 of a frame is very close to the maximum or reference amplitude, then it will be considered a "better" frame (high goodness). In actual implementation, FG at ith frame is defined in equation 4.8 as follows:

$$FG(i) = 20\log(MAXR0/R0(i)) \tag{4.8}$$

where MAXR0 is the maximum of R0 in an entire vocalic segment. Note that the lower FG is, the better frame goodness is. The values of FG at the vowel nucleus are often less than 5 while the FG reaches as high as

15 at the end of an utterance. That is, the amplitudes at these end-utterance frames are down 15 dB from their peak.

The main purpose of introducing FG in the SPG algorithm is to restrict the use of the "continuity rule" to those frames whose FG values are very high. In the past, one of the most powerful and frequently used rule applied to formant extraction has been the "continuity rule". It assumes that formant contours can not change too rapidly within a short time period, say 10 msec. Under this assumption, if formant extraction algorithms cannot assign the formant values at a current frame, then the programs simply use the formant values of a previous frame. Most of the false sudden rises or falls on formant contours can be smoothed out, based on the continuity rule. However, under this rule, an error made on a single frame may propagate through all the following frames. Additionally, we have found that formant values can change ligitimately as much as 700 Hz in less than 5 msec as in the case of the F3 transition between /IH/ and /R/ in the word "ZERO". Such rapid transition should be allowed by an adequate formant tracker.

Generally speaking, the SPG algorithm maps three sets of peaks (LMS, DPS and SDGD) to sensory formants using different strategies based on whether a current frame is a nasal frame or not. For non-nasal frames, the SPG algorithm starts from LMS peaks because they often represent formants (it is very rare to have spurious peaks on LMS) and searches for DPS peaks when it thinks that LMS does not have enough peaks. It checks SDGD peaks only if it fails to find enough peaks on DPS. Upon the completion of peak-checking, the SPG algorithm decides

the sensory formant assignments at a current frame based on either the newly computed values or on those of the previous frame. The SPG algorithm is executed at each frame of a strongest glottal-source (SGS) segment to generate a raw sensory path.

The following notations are used in the SPG algorithm both for non-nasal and nasal frames:

MDIST(i):
the MDIST computed at ith spectral peak;

LMS(i): (in dB)
the log magnitude of LPC spectrum at ith components;

DPS(i): (in angular-frequency)
the derivative of phase spectrum at ith component;

SDGD(i):
the second derivative of group delay at ith components;

FG(i):
the goodness of ith frame;

F(i): (in Hz)
the frequency location of ith spectral component;
the example assignments of F are listed below:

F(1)=0
F(2)=19.57
F(3)=39.14
...
F(51)=998.
...
F(512)=10,000.

F3_limit:
the maximum value of F3 and F3_limit=2400+6MSR.

V1:
the first frame of an SGS segment;

AVGPITCH:
the average of pitch values in a SGS segment;

SF1L, SF1H, SF2 and SF3:
the sensory formants at ith frame unless indexed otherwise.

The SPG algorithm (non-nasal frame):

1. Find all the peaks in LMS

   $L_1, L_2, L_3, \ldots, L_n$

   where L_i points to ith peak and has satisfied the following conditions:

   $F(L_i) <= F3_limit \qquad i=1,2, \ldots, n$

2. Delete any peak i if

   $LMS(L_i) < LMS(L_{i+1}) - 20$

   the resulting peak sequence is $R_1, R_2, \ldots, R_m$.

3. If m=3 then SF1L=SF1H=$F(R_1)$, SF2=$F(R_2)$, and SF3=$F(R_3)$.
   GOTO 14

4. If m >= 4 then do the following:

   4.1 find all the peaks in DPS

   $D_1, D_2, \ldots, D_p$

   4.2 sort the peaks in the descending order based on their DPS values, i.e., the sorted peaks $S_1, S_2, \ldots, S_p$ satisfy the following conditions:

   $DPS(S_1) > DPS(S_2) > \ldots > DPS(S_p)$

   4.3 Sort peak S1, S2 and S3 according to their frequency locations, resulting sequence is DS_1, DS_2, and DS_3 so that $F(DS_1) < F(DS_2) < F(DS_3)$.

   4.4 SF1L=SF1H=$F(DS_1)$, SF2=$F(DS_2)$, SF3=$F(DS_3)$

   4.5 GOTO 14

5. Compute MDIST(1) at D_1.
   If MDIST(1) < 0.8 then SF1L=SF1H=$F(D_1)$, GOTO 6
   Else If current frame is V1
   Then SF1L=SF1H=500
   Else SF1L=SF1H=SF1L(i-1)
   End if
   D2=D_1, Goto 7
   End if

6. Find the first peak after D_1, called D_2.

7. Compute MDIST(2) at D_2.
   If 0.56 < MDIST(2) < 1.2 and SF2-SF2(i-1) < 1000
   Then SF2=F(D_2)
   Else SF2="undefined", D_3=D_2
   End if.

8. If there is no peak after D_2
   Then SF3 = SF3(i-1), Goto 14;
   Else find the first peak after D_2, call it D_3.

9. SF3=F(D_3).
   Compute MDIST(3) at D_3.

10. If SF3 - SF3(i-1) > 400 and i > V1
    Then find the first two peaks after D_2 on SDGD, call them
         S1 and S2 (i.e., S1, S2 > D_2);
         Compute MDIST(3) at S2;
         If 0.86 < MDIST(3) < 1.4 and F(S2) < F3_limit
         Then SF3=F(S2), Goto 14 End if;
    End if.

11. If MDIST(3) < 0.86 and it is the first time to appear
    Then find the first peak after D_3, call it D_3.
         If there is no peak found
         Then Goto 8
         Else repeat step 9 - 11.
         End if;
    End if.

12. If |SF2-SF2(i-1)| > 200 then do the following:

    12.1 If SF2 matches a peak on LMS (difference < 100)
         Then Goto 13;

    12.2 If SF3 matches peak i on LMS (difference < 100)
         and SF2(i-1) matches peak i-1 on LMS (difference < 100)
         Then SF2 resulted from a spurious peak therefore
              reassign SF2 to the value of SF3;
              Goto 13;

    12.3 SF2 = SF2(i-1);

13. If |SF3-SF3(i-1)| > 300 and FG < 15 then do the following:

```
13.1  If |SF3-SF3(i-1)| > 600 Then
         If SF2(i-1) < 1800+AVGPITCH*2 and
            SF3(i-1) < 2200+AVGPITCH*3 and
            SF3/SF2 < 13.6
         Then
           (1) Search a peak (P3) on SDGD between SF2+50 and SF3;
           (2) If found and SF3(i-1)-F(P3) < 500
                 Then  SF3=F(P3), Goto 14;
           (3) Goto 13.2
         End if;
      End if;

13.2  If SF3 matches a peak i on LMS (difference < 150)
      Then Goto 14
      Else
          If  FG > 10 Then SF3 = SF3(i-1)
      End if

14. End
```

The sensory spectra in a nasal frame are relatively easy to compute compared to those in a non-nasal frame. Generally, there are only two major energy concentrations in the low-frequency region (below 3000 Hz) when LPC spectrum is plotted in dB vs. frequency for a nasal frame. In such a case there is only one wide spectral peak between 200 Hz and 1500 Hz. Therefore, we will estimate SF1L and SF1H based on this spectral peak. SF2 and SF3 are estimated from the second spectral peak, which is usually located around 2000 Hz. The SPG algorithm for computing sensory formants at nasal frames is described below:

The SPG algorithm (nasal-frame).

1. SF1L is the strongest peak on DPS whose frequency is higher Than MSR but below 300 Hz. Otherwise SF1L=MSR;

2. If there is no peak on DPS below 300 Hz and above MSR, Then SF1L takes the previous value.

3. If there is a peak on LMS in (500,1500) region,
   Then take it as SF1H; otherwise select right-edge of P1
   (20 dB down) as SF1H.

4. If there are two peaks on LMS in (1500, F3_limit)
   Then take them as SF2 and SF3, respectively.
   Otherwise, do the same for DPS.

5. If there is only one peak on LMS in 1500, F3_limit)
   Then take it as SF3 and take the left-edge of this peak
   (3 dB down) as SF2.

6. If the algorithm fails to compute sensory formants at above steps,
   Then take the previous values for the current frame.

4.3.5 Performance of the SPG Algorithm

In our project, we are interested only in the performance of any formant-tracking algorithm tested on real speech, not on synthetic vowels. However, it is impossible to determine the accuracy of any formant-tracking algorithm for real speech, since there is no way to ascertain the value of the formants in natural speech. The only realistic approach is to ask experts to manually label formants frame by frame after (1) learning the meaning of the utterance; and (2) examining the LPC spectra plotted in dB vs. frequency. Then formant contours generated by an algorithm can then be compared with those generated by humans assuming that the same definitions for formants are used by both methods.

However, it is extremely time-consuming to manually label formants on spectrum plots because there are about 100 glottal-source frames on average in an utterance. In TRAIN and TEST1 there are 88 tokens, i.e., more than 8000 spectrum plots have to be examined manually to record more than 3200 formant contours (four contours per utterance). Based on this consideration, we have chosen only 20 tokens in TRAIN, one for each

talker (one male and one female) recorded from the words "ZERO" through "NINE". Since the decision on whether a frame belongs to an SGS segment is often different for a human and a SWIS classification algorithm, two different SGS segments result from the same utterance, that is, they start at different frames and have different lengths. When such a case is encountered, they are aligned so that the difference between two sets of formant contours can be compared.

Formant tracking errors are defined as described below. The absolute difference between two sensory formants is computed. For example, if the SPG algorithm computed SF2 at ith frame is 2300 Hz and the human labeled SF2 at the same frame is 2100 Hz then the difference at ith frame is 200 Hz. Unfortunately, the absolute difference of the same formant computed by two methods may have different meanings depending on the formant from which they are calculated. For example, a 300 Hz error may be considered a minor error for the third sensory formant while the same number will indicate significant discrepancy between the two methods for SF1L or SF1H. To capture the significance of an absolute difference derived from two methods, the relative error is used. Here, there is a problem of establishing a reference value for a given formant. For simplification we have chosen the average of the formant values computed by the two methods as reference value. As such, from the previous example, the relative errors for SF2 at ith frame will be about 9%, computed from 200/((2300+2100)/2). Finally, the mean of relative errors is calculated for a formant over all the frames in a word.

Figure 4.4 shows the means of relative errors for each formant computed from one utterance per vocabulary word per talker. As seen in Figure 4.4, most of the means of relative errors are in the range of from 4% to 6%. Note that the SPG does better for females than males in the low frequency region of SF1L and SF1H. In the higher frequency regions, the performance of SPG is more variable with both the vocabulary words and the sex of the talkers. For example, the mean relative errors in percent is twice as high for females as for males for the word "TWO" (at SF2 chart) and also for the word "EIGHT" (at SF3 chart). The more detailed SPG error listing can be found in Appendix 10.3, where minimum and maximum absolute differences between the two methods are listed for each formant.

Also, we wish to show some sensory formant contours that have been generated by the computers so that the readers might make their own judgments on the performance of the SPG algorithms. In Appendix 10.4, we provide the 22 sets of sensory formant contours generated by SWIS from the tokens in TRAIN. For each vocabulary word in DIGIT, the two sets of sensory formant contours are plotted on a Hewlett-Packard 2623A graphic terminal, one for a male subject and one for a female subject. Note that these sensory formant contours are generated from the strongest glottal-source segments of the utterances. The waveforms corresponding to each SGS segment are also plotted under the sensory formant contours. In this way, the reader can easily relate the waveforms to individual segments on a set of sensory formant contours. Note that under each waveform plot there are two curves plotted along the time domain. The dotted line represents pitch values while the continuous curve is the unsmoothed signal amplitude.

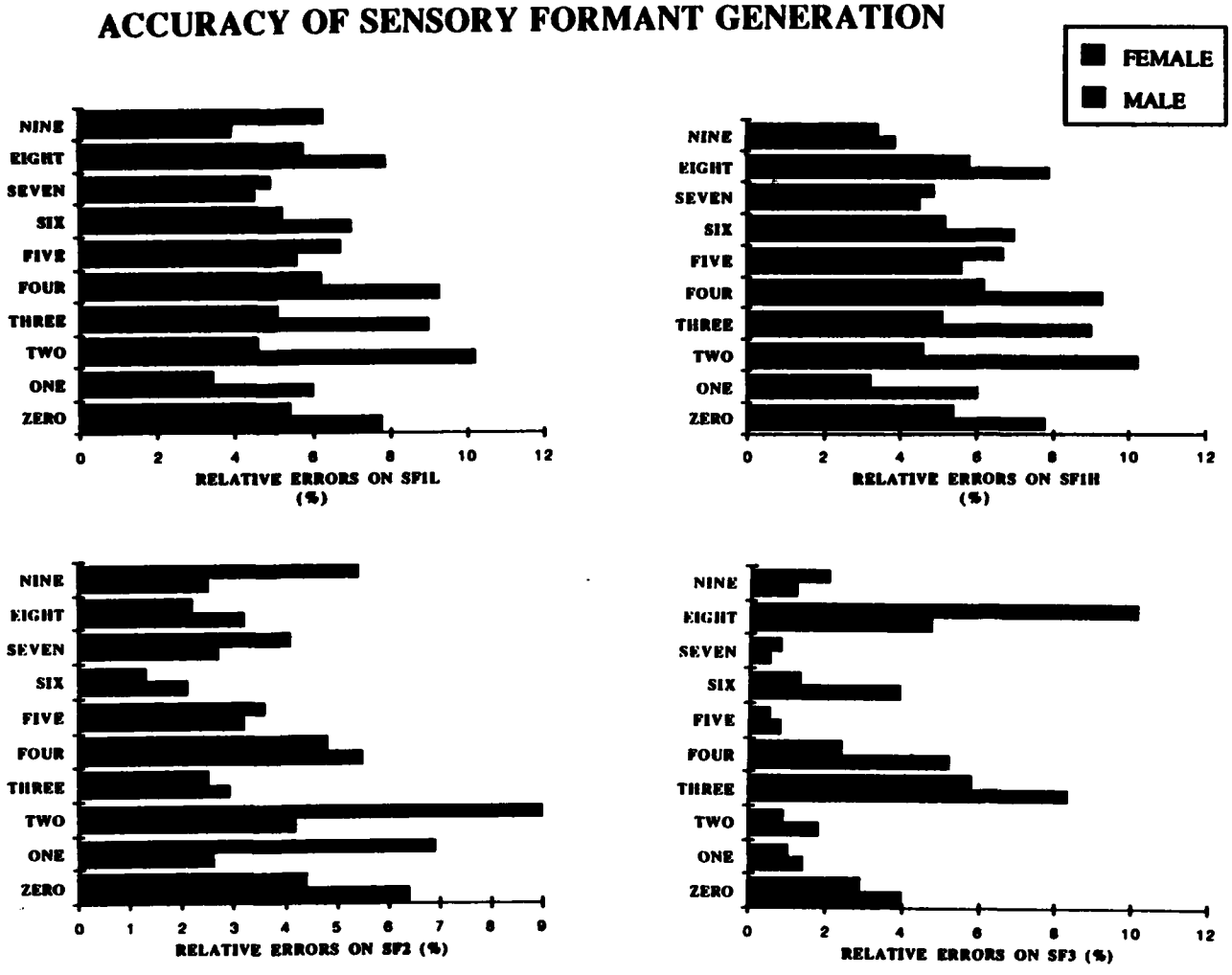


Figure 4.4 Mean error rates computed over the four sensory formant contours (SF1L, SF1H, SF2 and SF3).

5. PHONETIC DECODER

5.1 INTRODUCTION

In this chapter we present a detailed discussion of the Phonetic Decoder (PD) developed in the SWIS project. As its name suggests, the main function of the PD is to derive a sequence of phonetic codes from a sensory path which is in turn generated from an utterance by the SWIS front-end system.

The PD consists of two major components, namely, the perceptual path generator (PPG) and the phonetic parser (PP). The PPG is implemented as a transformation system which takes a smoothed sensory path derived from an SGS segment as input, and outputs a corresponding perceptual path. Note that only the SGS segment in an utterance is processed by the PPG and, therefore, there is no perceptual path for all the other segments.

The structure of the PPG and its design are described in section 5.2. The phonetically-encoded auditory-perceptual map is described in section 5.3. Based on this map the PP is developed to parse a continuous perceptual path, to decide how many phonetic elements are contained in the path, and what they are. The recognition process for pure vowel (non-diphthongs) based on the PP is described in section 5.4. The recognition of the two diphthongs relevant to this project is based on a different strategy which is discussed in section 5.5. In section 5.6, we provide a tentative phonetic dictionary for DIGIT, which is based on the results of running SWIS for the test tokens in TRAIN and

TEST1. In the last section of this chapter, we will discuss a simple lexicon-accessing scheme used in SWIS for word generation.

5.2 SENSORY-PERCEPTUAL TRANSFORMATION

In this section, we present a sensory-perceptual transformation system used in the phonetic decoder (PD) of SWIS. Generally speaking, this transformation system translates a smoothed sensory path into a perceptual path in a similar to that described in APT (see Appendix A). However, the transformation equation used in SWIS is much simpler than the one proposed in APT because the sensory paths input to this system contain only vocalic segments within an utterance, while the one proposed in APT is designed to handle not only the transition segments between vocalic segments and burst segments, but also the burst segment itself. Furthermore, the transformation of SWIS is defined in the linear frequency domain while that of APT is in the logarithmic frequency domain.

The transformation system is essentially a second-order resonator originally proposed by Klatt (86) for use in a formant-based synthesizer. The recursive discrete implementation for the second-order resonator is shown in equation 5.1, where X and Y represent input and output variables, respectively.

$$Y(nT) = A*X(nT) + B*Y(nT-T) + C*Y(nT-2T) \qquad (5.1)$$

where

$$C = -e^{(-2*PI*BW*T)}$$
$$B = 2*\cos(2*PI*F*T)*e^{(-PI*BW+T)}$$
$$A = 1-B-C$$
$$PI = 3.1415926$$
$$BW = 2*D*F$$

In the equation 5.1, T is the sampling period; D is the damping factor; and F is the center frequency of the resonator. Note that a current output value from the equation 5.1 depends on a current input value and two previous output values, that is, evaluation of the recursive equation requires the two initial valuess. In actual implementation we simply assume that the two initial output values are equal to the first input data value. This means that the transformation system always responds quickly to the input data in the first 6.4 msec.

The output of the equation is controlled by F and D. Generally speaking, the lower the value of center frequency (F), the greater the value of damping factor (D), the smoother the value of the output curve becomes. The relationship between the input to the transformation system and the output from the system can be best illustrated by plotting a smoothed sensory formant (input to the transformation system) and a "perceptual" formant (output from the transformation system) in the time domain as shown in Figure 5.1. In both the upper plot and the lower plot of Figure 5.1, the curves drawn in darken line are related to the same smoothed SF3 (from a male utterance "THREE"). However, the output variables, i.e., the resulting perceptual formants output from the transformation system (drawn in the dash lines), appear quite different for the two different transfer functions (the center frequencies are different). Note that the "overshoot" of the output values, shown in the upper plot of Figure 5.1, becomes visible in the region where the input variable undergoes relatively rapid changes if a high center frequency (30 Hz) is used in the transfer function. The transformation system is not as sensitive to changes in D as it is to

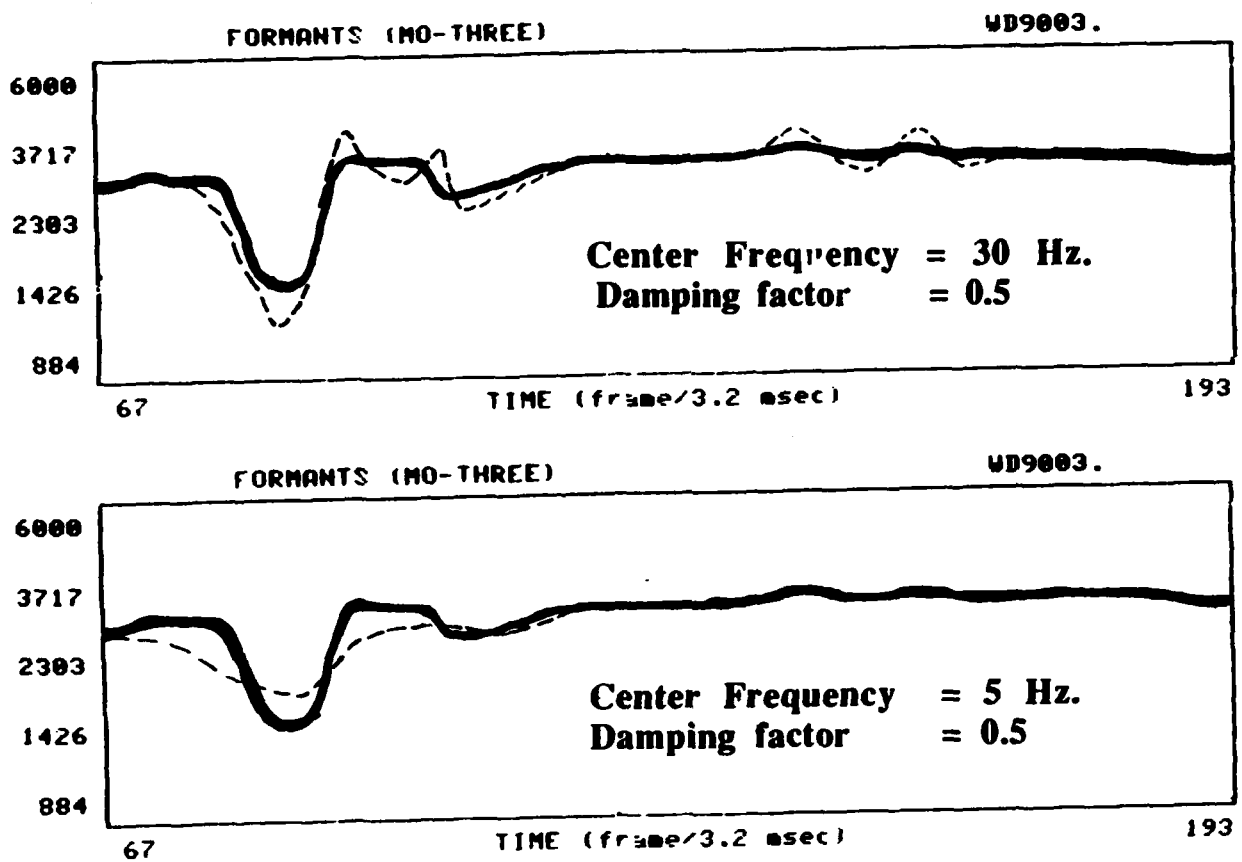


Figure 5.1 Effects of using the two different center frequency in the sensory-perceptual transfer function for a same input variable (smoothed SF3 in darken line), where the output from the function is ploted in dash line.

changes in F. However, we have found that the damping factor (D) should be set close to 0.6 for a satisfactory result.

After experimenting with many combinations of F and D with many sensory formant contours derived from the training database TRAIN, we arrived at the following values for F and D.

$$F = 5 \quad (Hz.) \tag{5.2}$$
$$D = 0.5 \tag{5.3}$$

When these two values were used in SWIS to test the tokens in TRAIN, the resulting 44 perceptual paths for TRAIN had no overshoot at all. After that, we tested the same transformation system with the initial testing database TEST1, and the results were equally satisfactory.

As we discussed earlier, a sensory path SP(X,Y,Z) is defined as a curve in a 3-dimensional Auditory-Perceptual Space (APS), where X, Y, and Z are coordinates along each axis of APS. Thus, we will have the three input vectors to the transformation system as defined in the following equations.

$$X = \text{Log}(SF3/SF2) \tag{5.4}$$
$$Y = \text{Log}(SF1L/SR) \tag{5.5}$$
$$Z = \text{Log}(SF2/SF1H) \tag{5.6}$$

The following iteration sequences show an example of how a sensory ratio (the ratio of SF3 to SF2 is used in the example) is transformed into a perceptual ratio for an N-frame glottal-source segment using the sensory-perceptual transfer function defined in equation 5.1.

```
X(i) = SF3(i)/SF2(i)                    i=1,2, ..., n
Y(1) = X(1)
Y(2) = Y(2)
...
Y(i) = A*X(i) + B*Y(i-1) + C*Y(i-2)
...
Y(n) = A*X(n) + B*Y(n-1) + C*Y(n-2)
```

The A, B, and C are the transformation coefficients calculated from F and D (see equation 5.1). The resulting perceptual ratio is contained in Y. Thus, the same transformation system is executed three times for each sensory ratio variable.

In the next section, we will describe how a phonetically-encoded auditory-perceptual map is developed, based on the perceptual paths computed from TRAIN and TEST1.

5.3 PHONETIC-ENCODED AUDITORY-PERCEPTUAL MAP

The concept of a phonetically-encoded auditory-perceptual map (refered to as a P-map hereafter) is similar to an ordinary geographic map. For example, once a path is decided, one can determine which cities/counties will be passed on the way by simply tracing the path on a map. The basic idea of using a P-map is as follows. Assume that a perceptual path, traveling in APS, contains relevant phonetic information. Assume also that there is a way to establish (on a P-map in APS) zones or boundaries for phonemes. Then, once a path is specified we can simply trace it on the P-map to see which areas (one for each phoneme) it passes through. The names of these passed-through areas will then correspond to the possible phonetic events occurring along the path. Up to this point, we have answered the question of how a perceptual path can be generated from an utterance. We now must address other issues in the context of phonetic recognition, that is, how do we

find appropriate areas or subspaces in APS for all the phonemes relevant to this study ? Of course, once these areas or subspaces are established, the design of a P-map becomes obvious.

First, we list all the phonemes for which we plan to define areas on the P-map or on subspaces in APS. They are as follows:

/IY, IH, EH, AE, AA, AH, UH, AO, UW, W, R/

Among these eleven phonemes, two are consonants, i.e., /W/ and /R/. Except for these two consonants, Miller and his associates (88-89) have extensively studied the aforementioned phonemes in the context of their subspace specifications. In these studies, the basic approach to the problem was to manually place points in APS, based on either formant contours extracted from real speech or formant values published in the literature. In doing this, they have reverted to the more traditional method of speech research, wherein the investigator selects the appropriate segments in **a posterior** fashion, based on all the available knowledge including the sound of the segment.

Sensory reference values were calculated from the talker's pitch based on the formula proposed in APT. If a pitch value is not available but the sexes of the subjects are known then 155 Hz and 185 Hz are used as sensory reference values for men and women, respectively. If no information about the subjects is available then 168 Hz is used as the sensory reference. Each point is labeled based on its phonetic content. If a point is computed from an /IH/ segment, then it is labeled as a /IH/ point. After a sufficient number of points are placed in APS using the selection techniques described above, the points when grouped by

their labels start to form their own subspaces in APS. This phenomenon becomes even more visible when these points are transformed from their orignal space (APS) to a new space SLAB(X',Y' and Z') and then projected to the X'-Y' plane of SLAB. The SLAB space is a simple mathematical transformation from APS. The transformation is defined in the following equations:

$$X' = 0.70711(X-Y) \qquad (5.7)$$
$$Y' = 0.81622Z - 0.4081(X+Y) \qquad (5.8)$$
$$Z' = 0.57722(X+Y+Z) \qquad (5.9)$$

The detailed discussion about SLAB can be found in the Appendix A. Finally, the boundaries of these subspaces are manually drawn based on the X'-Y' projection, with the intention that there should be no overlap among them. Thus, each subspace appears as a 2-dimensional irregularly-shaped enclosed area, similar to a county or state boundary on a geographic map (they are called phonetic target zones in APT.) It is based on these target zones that our P-map is initially established.

Note that the P-map has only 11 target zones. This does not mean that only 11 phonemes can be represented in APS. In fact, there is a large-scale ongoing project headed by Miller to quantitatively specify target zones for all English consonants (90).

The P-map has been iteratively modified during the course of the SWIS project as we have incorporated data from more and more perceptual paths, first from the training database TRAIN and then from the test database TEST1. Two types of modifications are involved in the development of the P-map, namely, (1) expansion and (2) conflict resolution.

The modification process is as follows: We plot a perceptual path on the P-map. Knowing the meaning of that path, we can make a subjective judgment on whether it passes through all of the necessary target zones. If a particular segment on that path misses a target zone because it is outside of that zone **and** is not inside any of the other zones, then we expand that zone so that the segment will stay inside after the expansion. This type of modification generally does not introduce a "chain-reaction".

However, if a particular segment on a path misses a target zone because it enters an inapproriate target zone, then we must carefully resolve this "region conflict" or a chain-reaction will occur. For example, if we redraw the map by expanding some zones and shrinking others to fit that path, it may cause other paths to miss their target zones. In a sense, all these target zones compete with their neighbors in terms of their territories. The only way to solve these region conflicts is to use majority-rule. That is, every time the map is revised due to this kind of conflict, all the available perceptual paths are replotted to see how many misses there are. Let's use an example to explain how this is done in SWIS. Consider two phonetic target zones, A and B, that are adjacent. Suppose that there are ten A segments falling into zone B and five B segments falling into zone A. Thus, we have 15 misses based on the current version of the P-map. Now, we make a revision on the P-map, resulting in a situation where only three A segments are still inside zone B, but six B segments are inside zone A. Although one more B segment has missed its target zone due to the new version of the P-map (because the B zone has been shrunk), the total

number of misses has been reduced from 15 to 9. Thus, this new version will be retained as the current version of the P-map. Otherwise, the revision would be undone, i.e., we would retain the previous version as the current version of the P-map. Almost all the modifications based on the conflict resolution scheme occurred at the boundaries between two phonetic target zones. So, conflict resolution is very easy to carry out by redrawing boundaries in such a way that the majority of misplaced segments will be moved into appropriate pnonetic target zones. Because there are only 44 perceptual paths in TRAIN, the iteration process for conflict resolution is quite fast. After that, we apply the same techniques for 44 perceptual paths in TEST1. After the final modifications, all the phonetic segments on these 88 perceptual paths have been placed in appropriate phonetic target zones.

The final version of the P-map is shown in Figure 5.2. All the target zones except for /R/ have their Z'-coordinates values falling between 0.6 and 0.8 while /R/ has as its Z'-coordinate value less than 0.6 as shown in the sideview of the P-map in Figure 5.2 The /R/ target zone in the P-map is substantially larger than the original one specified in APT because the boundaries of the /R/ transition segments, both those moving away from and into /R/, are very difficult to locate. To simplify our search algorithm, we define that a segment is in the /R/ target zone if it falls below the vowel slab, i.e., its Z'-coordinate value is less than 0.6. In this way, the resulting /R/ target zone covers from -0.2 to 0.4 along the X'-axis and from -0.2 to 0.5 along the Y'-axis as shown in the front-view of the P-map in Figure 5.2.

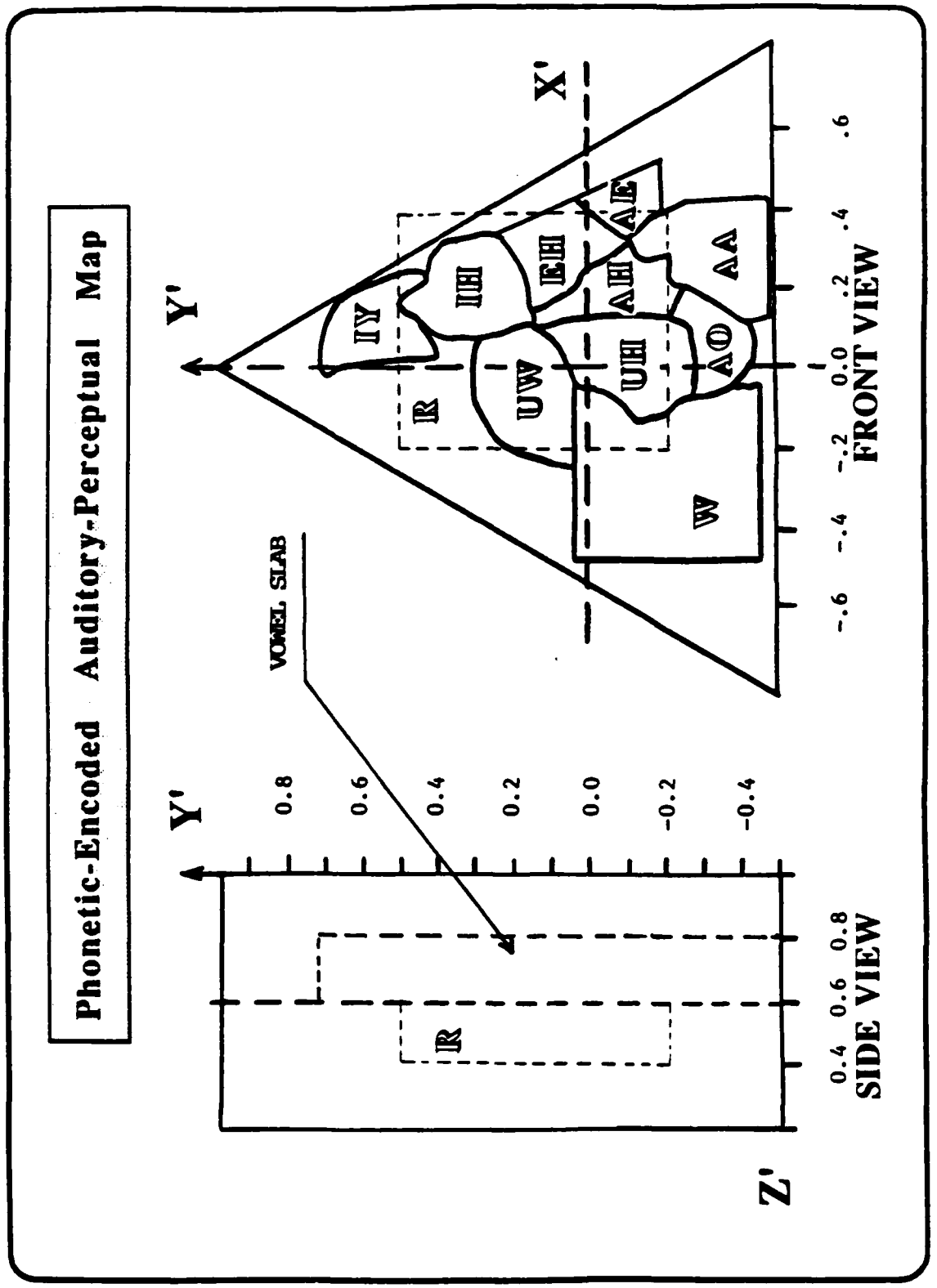


Figure 5.2 Sketch of the Phonetic-Encoded Auditory-Perceptual Map

The P-map also differs from the original target zone definitions in APT (not shown), in that the target zone for the diphthong /OW/ has been completely deleted. There are two reasons why we made this change. First, retaining /OW/ caused too many "false alarms", i.e., many non-/OW/ segments passed through the area. For example, almost all the perceptual paths generated from the words "FOUR" and "ONE" passed the target zone of /OW/ proposed in APT. Furthermore, the speed of the movements along these paths is comparable to these along the perceptual paths generated from /ZERO/, which does in fact contains the sound /OW/. Second, we believe that diphthongs should be represented in "directional pipes" in APS rather than defined as passages over a mutually exclusive area in the P-map. The representation of diphthongs used in SWIS will be discussed in section 5.5.

5.4 CRITICAL SEGMENT DETECTION

At the early stages of the development for the Phonetic Parser (PP), we employed a very simple tracing technique to determine how many phonetic segments are in a perceptual path and what they are. With the tracing technique, we simply count how many target zones a perceptual path passes and record the sequence of labels of these target zones as the sequence of phonetic codes. If a target zone has been passed twice, then it will be counted twice and there will be two occurrences of that phonetic label in the final phonetic code sequence generated by PP.

The problem with this tracing technique is that there are too many phonetic codes generated from each perceptual path. However, it is still possible to construct a very complex phonetic dictionary based on these phonetic code sequences derived from the training databases TRAIN

and TEST1, because there are only 11 words in the vocabulary. But it is highly questionable that this method can be generalized to construct a large-size phonetic dictionary.

To overcome the problem of more phonetic codes being associated with a perceptual path than needed, we developed a "critical segment" detection technique to assist PP. Then, the previously-described tracing method is applied only along these critical segments. Hopefully, this accurately reduces the total number of phonetic target zones encountered on a perceptual path. First, we have to decide how many critical segments should be identified, given a perceptual path. As we discussed earlier, the main goal of the Phonetic Decoder in SWIS is to recognize the **most-stressed** syllable in a word, and we know that there are at most three vocalic phonemes in any one syllable of our set. Thus, it becomes clear that only three critical segments are needed, and they should be located in the region of the most-stressed syllable. Here, the working assumption is that each of these three critical segments will pass a set of phonetic target zones, one of which is the phoneme constituting the nucleus of the most-stressed syllable.

Now, we present the critical segment detection (CSD) algorithm used in PP.

The Critical Segment Detection Algorithms

1. Find an anchor frame **A** in a perceptual path; assume that most-stressed syllable segment is centered on the frame **A**.

2. Find the first critical segment, called Pre-Critical, in the region to the left of **A**.

3. Find the second critical segment, called Post-Critical, in the region following **A**.

4. Find the third critical segment, called Tail-Critical, in the region following the Post-Critical segment.

Clearly, once we locate three critical segments, we simply apply the tracing technique described above to each segment. This is done to record the phoneme labels associated with each target zone encountered by the segment. Generally speaking, there are at least three phonetic codes generated by PP for a given perceptual path, assuming that each critical segment encounters at least one phonetic target zone. Occasionally, Pre-Critical and Tail-Critical segments from a perceptual path fall completely outside of any phonetic target zone. In such cases, PP may output only one phonetic code (from Post-Critical segment) for that perceptual path (it often occurs in the word "TWO"). Usually, each critical segment passes more than one phonetic target zone (two on average). In the next four sections, we will detail each of four steps of the CSD algorithm. The following notations are used in the CSD algorithm:

V(i):
The velocity value at ith frame (see section 3.2.7);

A(i):
The acceleration value at ith frame (see section 3.2.7);

SI(i):
The segmentation index at ith frame (see section 3.2.8);

V1:
The first frame in a perceptual path;

V2:
The last frame in a perceptual path.

5.4.1 Anchor Frame

The purpose of locating an anchor frame (A) is to make an educated guess as to where the most stressed syllable (MSS) would be on a given perceptual path. By examining the amplitude contour (R0), we often see the strong correlation between an MSS and the highest peak on R0. That is, the highest peak frame (PEAKFN) is almost always located inside an MSS. Based on this observation, we simply define A as PEAKFN, where R0 reaches the highest value in an entire perceptual path.

During the initial testing with the tokens in TRAIN, all the anchor frames found in this way corresponded to their appropriate MSSs. However, when the same anchor selection program was tested in TEST1, several perceptual paths were mislabeled for their anchor frames. The reason was that in these cases there were two very high peaks on R0 contour, and the second one was the highest and therefore chosen as A. Actually, the first high peak corresponded to MSS, not the second. The examples for such cases are the amplitude contours extracted from the words "ZERO" and "SEVEN." Because our previous rule worked for most tokens in both TRAIN and TEST1, there only slight modification was made on the selection program so that the first strong peak is chosen as A if the ratio of this peak value to the highest peak value is above 95%. After this modification, all the anchor frames selected corresponded to their MSSs.

5.4.2 Pre-Critical Segment

Once an anchor frame (A) is selected, a Pre-Critical segment is found by a backwards search from A to V1. At the end of searching, a

Pre-Critical segment (P1,P2) is located where its starting frame number P1 >= V1 and its ending frame number P2 is < A.

Generally speaking, a Pre-Critical segment always covers the first phoneme in CVC or CV syllable if C is in a vocalic segment. For example, /R/ in the CV syllable /R+IY/ (in the word "THREE") is often identified through the analysis of a Pre-Critical segment. In case the first phoneme in the CV syllable is outside of a vocalic segment, such as /Z/ in "ZERO", we hope that an underlying Pre-Critical segment will at least pass the target zone /IY/, i.e., to identify the vowel segment for that CV syllable. The detailed procedure for the selection of P1 and P2 is as follows.

Step 1: Search backwards from A to find the first frame E where SI(E) > 0.0; If such an E does not exist in the region (V1,A), then find the lowest velocity point in the region (frame E);

Step 2: Set P1 and P2 equal to E;

Step 3: Move P1 to left until either V(P1) < 4 log-unit/sec or P1 = V1;

Step 4: Move P2 to right until either V(P2) < 4 log-unit/sec or P2 = A;

5.4.3 Post-Critical Segment

The selection of a Post-Critical segment is more complicated than the selection of a Pre-Critical segment. In an actual experiment, we notice that once perceptual paths are divided by frame A into the two segments, the first one which precedes A is usually shorter than the one following A. This means that we have more frames with which to work to locate a Post-Critical segment. In addition, there will be more non-zero

SI points in the second segment, which causes difficulties in locating the boundaries of a Post-Critical segment.

Generally speaking, we wish to find a Post-Critical segment as close to A as possible. If this Post-Critical segment does not cover the second phoneme in a CVC syllable, we still have a chance of finding it through a Tail-Critical segment. On the other hand, if a post-critical segment contains only the last phoneme in a CVC syllable there is no way to recover this mistake by locating a Tail-critical segment. The following procedure describes how a Post-Critical segment (P3,P4) is selected from a region starting at frame A and ending at frame V2, which is the last frame of a perceptual path.

Step 1: Search backwards from A up to 63 frames to find the first frame E where SI(E) > 0.0;
If such an E does not exist in the region (A,V2), then find the lowest velocity point in the region (frame E);

Step 2: Set P3 and P4 equal to E;

Step 3: Find frame VMIN where

$$V(VMIN) < V(i) \quad i=A,A+1, \ldots, V2$$

Step 4: If SI(E) > 0.0 then
If VMIN < E then move P3 to left until either

V(P3) > 4 log-unit/sec or
P3 = A

Else
Move P3 to right until either

V(P3) > 4 log-unit/sec or
P3 = V2

End If
End if;

```
Step 5:  If SI(E) > 0.0 then

            find the next non-zero SI(El);
            and then set P4 equal to El;

         If El does not exist then

            P4 = min(P3+32,V2);
            Goto Step 7;

         End if;


Step 6:  If El is found in Step 5 then
            Move P4 to right until either

              A(P4) < 0  and  A(P4+1) > 0  or
              { acceleration undergoes "-+" pattern }

              P4 > 4 log-unit/sec  or
              P4 >= V2
         End If;


Step 7:  If P4-P3 < 30 {Post-Critical segment is too short}
         Then
            Move P3 to left until either

               A(P3) > 0  or
               P4-P3 > 30
         End If;
```

5.4.4 Tail-Critical Segment

If a Pre-Critical segment and a Post-Critical segment together span more than half of a perceptual path, it is reasonable to believe that they should cover the underlying stressed vowel because they are located in the region where the most stressed syllable might occur. However, if both Pre-Critical and Post-Critical segments are found in the region close to the beginning of a perceptual path, say, all in the first half of the path, then it is quite possible for them to fail to cover the underlying stressed vowel. In other words, a perceptual path might

reach its most stressed vowel target zone in the second half of its entire journey. We found that this is often true for quite a number of the perceptual paths generated by the word "TWO", /T+UW/, where the target zones /UW/ are encountered at the end of these paths.

An obvious method of solving this problem is to locate one more segment following the Post-Critical segment with the hope that this "Tail-Critical" segment will capture target zones that would be missed otherwise. In order to avoid generating too many phonetic codes, a Tail-Critical segment is searched only if both Pre-Critical and Post-Critical segments are found in the first six-tenths of a perceptual path.

The basic strategy used in the Tail-Critical segment searching algorithm is to locate the first significant phonetic event following a Post-Critical segment. This generally means finding the first non-zero SI(i) point. If, however, the values of segmentation index (SI) in all the following frames are zero values, the lowest velocity point in the region between the last frame of a Post-Critical segment and the end of a path is selected. The following procedure describes how a Tail-Critical segment (P5,P6) is selected based on our discussion above.

```
Step 1:  If P4 > 0.6(V2-V1)+V1  Then Goto Step 5;

Step 2:  If there is a frame E such that
            SI(E) > 0  and  E > P4
         Then
            P5 = min( V(i), P4 < i < E)
            P6 = E
            Goto Step 4
         End if

Step 3:  P5 = min( V(i), P4+10 < i < V2)
         P6 = P5
```

Step 4: Move P5 to left until either

P6 - P5 > 30 or
P5 <= P4

Move P6 to right until either

P6 - P5 > 30 or
P6 >= V2

Step 5: If P6 - P5 <= 5 Then 0 -> P5, P6

The length threshold of 30 frames (96 msec) used in Step 4 is to make sure that a Tail-Critical segment is long enough to cover an important phonetic event, if there is any, occurring in the second half of a path. On the other hand, if the resulting Tail-Critical segment is too short (5 frames, i.e., less than 16 msec), it is ignored by setting the starting and ending pointers P5 and P6 to zero so that no phonetic codes are generated later. Both thresholds are determined based on the analysis of the perceptual paths generated from the training databases TRAIN and TEST1.

5.4.5 Phonetic Code Generation

As we mentioned in section 5.4.3, each critical segment passes two phonetic target zones on average. That is, on average a total of six phonetic labels are recorded for a perceptual path. In addition, these six phonetic labels are not necessarily different, i.e., there may be some repetition in the sequences. This fact is often due to the same target zones being passed by two adjacent critical segments. For example, a Pre-Critical segment passes the target zones /IY/ and /IH/ and a Post-Critical segment passes the target zones /IH/ and /R/. Then, the following phonetic labels are recorded for both segments:

[IY + IH + IH + R]

In this case, we will assume that there are no other phonemes between the two occurrences of /IH/, that is, it is impossible for a perceptual path to leave target zone /IH/ and then enter some other target zones and then come back to target zone /IH/ without leaving any trace on the path. Remember that a Post-Critical segment is centered around the **first** significant phonetic event following a Pre-Critical segment. Based on this assumption, we merge all the adjacent phonetic labels recorded from critical segments. As a result, the phonetic label sequence in the above example will becomes [IY + IH + R].

After the merge operation, we find that there are still too many phonetic codes for most perceptual paths. For example, there should be at most three phonemes for a given perceptual path because all the perceptual paths generated in SWIS are from CV, or VC, or CVC syllables (if our classification program works properly). To further reduce the total number of phonetic codes generated for a perceptual path, two rules are used in the elimination processes, namely, the three-to-two rule and the two-to-one rule.

In the three-to-two rule, two phonetic codes are selected from the three phonetic codes generated within a critical segment. The selection method is based on the application of "majority-rule" to the number of frames each code represents. Consider that a critical segment passes the three target zones A, B, and C and the total numbers of the passed frames for each zone are 10, 15, and 20, respectively. Then we have to eliminate the phonetic label for the target zone A from the phonetic

code sequence generated from that critical segment because its duration is too short to compete with the other two target zones.

If the total number of frames in a critical segment is very small (a short segment) but it passes two target zones, then we assume that there is at most one phoneme associated with that segment, the other one being some kind of transition. The two-to-one rule is applied when the duration of a critical segment is less than 50 msec. and there are only two phonetic codes generated from the segment. The selection process is similar, to that discussed in the three-to-two rule.

Like many threshold problems, we sometimes run into difficulties deciding when to discard a phonetic code. For example, if the difference between the durations of the two passed target zones is very small, say, 20 frames vs. 22 frames, then which phonetic code should be retained if we are only allowed to keep one of them? When this time-related comparison does not help to make such a decision, we use a distance-related comparison in our decision making process. This is implemented as follows. If a critical segment has equal durations withing two target zones, then we calculate the distances it travels within each target zone. The one with the longer travel distance is assumed to be more important to our perception, therefore its phonetic code is chosen to represent the segment. If, in some extreme case, both time-related and distance-related measurements do not resolve this kind of competition (in a sense, both phonetic codes try to compete with each other to represent their critical segment), then we keep both because there is no way to recover the phonemes associated with that critical

segment if both phonetic codes generated from the segment have been discarded.

5.5 DIPHTHONG RECOGNITION

As we stated in the beginning of this chapter, the recognition of two diphthongs, /EY, AY/, is based on different strategy than those discussed in section 5.4. Unlike most pure vowels whose identities are best represented by a set of spectral patterns in their steady states, it is generally recognized that diphthongs do not have steady states and only display their identities through a dynamic shift of these spectra (90-96).

Perhaps the most extensive study on English diphthongs was conducted by Holbrook, which constitutes his Ph.d dissertation under the direction of Grant Fairbanks at the University of Illinois in the late 1950s (96). Figure 5.3 shows the classical drawing for six common English diphthongs based on the experiment mentioned above. All the six envelopes drawn in Figure 5.3 were derived through examining the formant contours extracted from 20 male talkers. One question to be asked regarding this mapping scheme is wether there will be more overlap among diphthongs if female speech is included. Another is how do we distinguish non-diphthong paths from diphthong paths ? Based on our data, we did find that there is slightly more overlapp between /EY/ and /AY/ once female speech is added into a testing database. But this is not a big problem for SWIS because there are only two diphthongs to be recognized. The main problem exists because many non-diphthong paths appear to be similar or identical to the diphthong paths. Also, we want

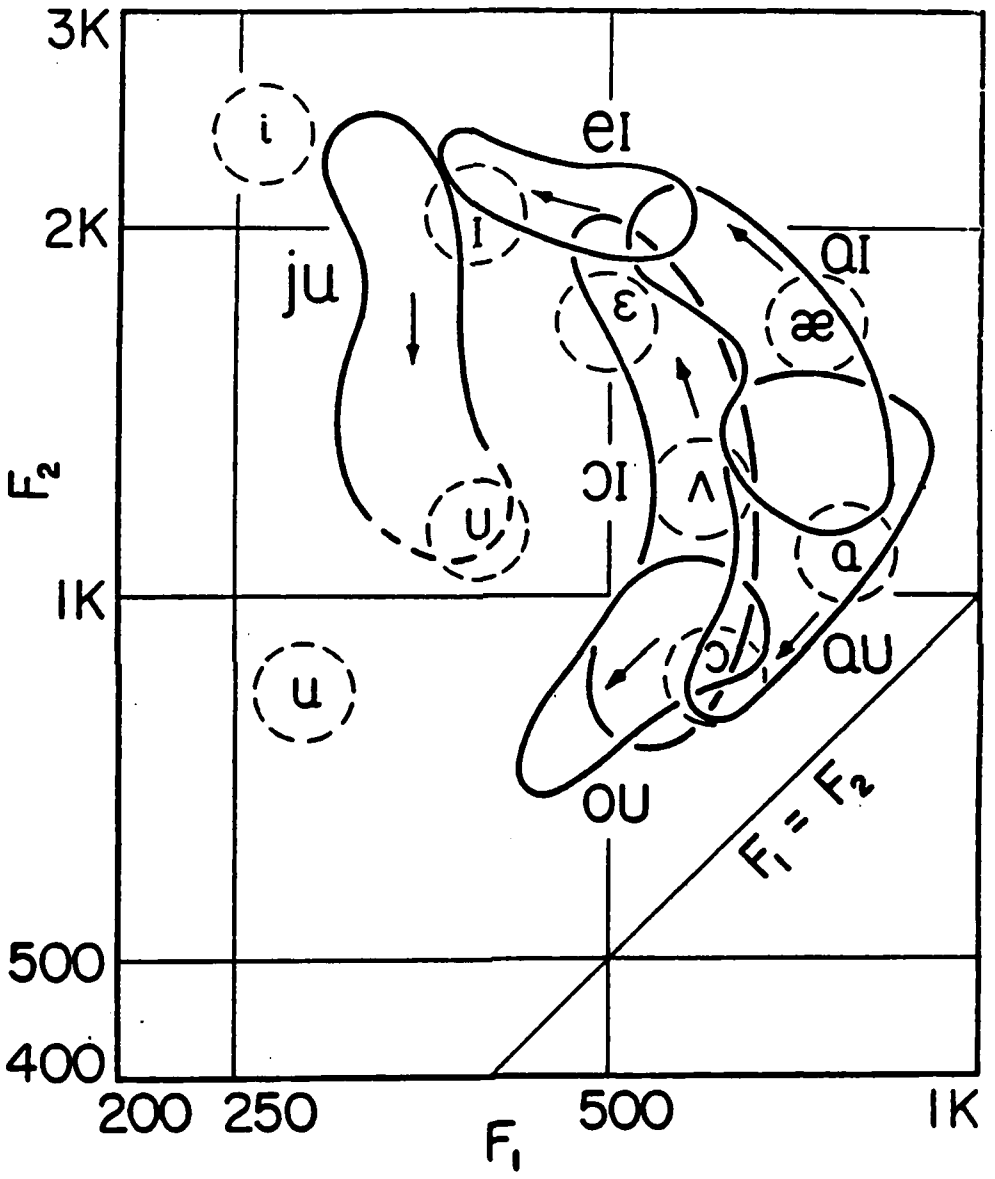


Figure 5.3 Time variations of the first two formants for six English diphthongs (after Holbrook and Fairbanks.)

to represent both pure vowels and diphthongs in a uniform fashion, i.e., we want to draw them on the same P-map as we did for nine pure vowels and two consonants in section 5.3 and not in the F1 by F2 plot of Figure 5.3. The following sections describe how this mapping scheme is achieved in SWIS.

Generally speaking, SWIS handles the problem of diphthong recognition using a method derived from consideration of a "pipe" technique wherein a directional three-dimensional pipe is specified in APS for each diphthong. A perceptual path is considered to contain a diphthong only if it (1) enters the pipe for that diphthong from a prespecified entry; (2) exits from the other end, and (3) satisfies a set of conditions during its journey inside the pipe. For example, it has to stay inside the pipe for a certain period of time. Possible pipes for /EY/ and /AY/ are suggested by the following consideration.

First, we simply plot all the perceptual paths generated from words "EIGHT", "FIVE", and "NINE" in TRAIN and TEST1 (12 paths.) Then, we manually cut the paths at the frames corresponding to the beginning and the ending of the underlying diphthongs. The resulting 12 diphthong segments are plotted on the partial P-map in Figure 5.4. It is easily seen that the starting locations of the /EY/ segments are located in the upper portion of the P-map (all Y' values are above 0.2), compared with ones starting at the very lower portion of the P-map for /AY/ segments (all Y' values are below 0.0.) And these segments can be enclosed in the two pipes as drawn in darkened lines. Once the two pipes are specified on the P-map in terms of X'- and Y'-coordinates, we eliminate, one by one, those irrelevant perceptual paths passing through either of

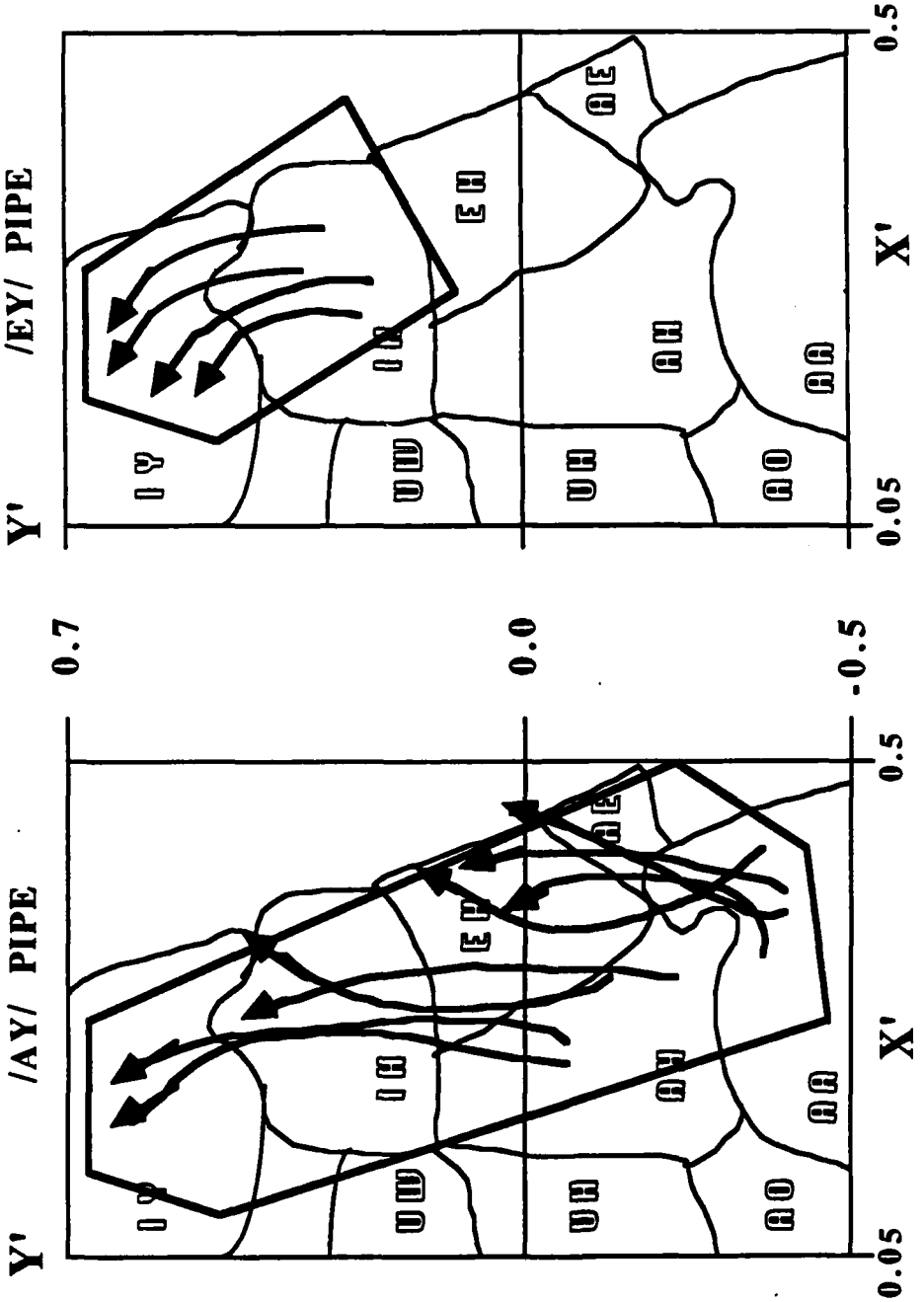


Figure 5.4 The perceptual paths of some diphthong segments are plotted on the P-map to illustrate the concepts of diphtong pipes.

the pipes by adding more restrictions. In this way, non-diphthong segments will be excluded. Of course, we have to make sure that we do include these 12 diphthong segments every time a restriction is added.

Recalling the concept of Y'-glide introduced in section 3.2.9, we note that concept of the upward pipe can be used to develop algorithms for recognizing which Y'-glides are associated with diphthongs. In this context the diphthong recognition algorithm was developed. Before the algorithm is presented, we list some important notations used in the algorithm, as below.

Assume that a Y'-glide starts at ith frame.

ANG(i):
The angle of the Y'-glide (see section 3.2.9 for details);

DUR(i):
The duration of the Y'-glide (see section 3.2.9);

LENGTH(i):
The length of the Y'-glide, defined as Y'(i+DUR(i))-Y'(i);

BASE(i):
The perceptual path location at ith frame;

Diphthong Recognition Algorithm

1. Find all the Y'-glides on a perceptual path;
2. A Y'-glide is a diphthong candidate if the following conditions are met:

   <1> ANG(i) > 45 or DUR(i) > 30;

   <2> 0.05 < Y'(i) < -0.05 if BASE(i) is inside of target zones for /EH/ or /AH/;

   Y'(i) < 0.4 if BASE(i) is inside of target zone /IH/;

   Y'(i) < -0.05 if BASE(i) is inside of target zone /AE/;

.1 < X'(i) and Y'(i) < -0.035 If BASE(i) is not inside of any target zones;

<3> LENGTH(i) > 0.1 or DUR(i) > 50;

<4> DUR(i)/(V2-V1) > 0.5 if DUR(i) < 40 or LENGTH(i) > 0.3;

<5> Y'-glide should not pass the following target zones:

/UH + AO + UW/

<6> Y'(i+DUR(i)) > 0.35 If Y'(i) > 0.05;

<7> LENGTH(i) > 0.15 if Y'(i) > 0.05

3. For any diphthong candidate found above:

it is /EY/ if Y'(i) > 0.05;

it is /AY/ if Y'(i) < -0.05;

Once a diphthong segment is identified from a perceptual path, all the previous labelings from critical segments are overridden by this diphthong label. We make no attempt to identify which consonant precedes or follows that diphthong. This is because phoneme /F/ and /V/ (in word "FIVE" : /F+AY+V/) are not covered in this project and the identification of [nasal][AY][nasal] sequence is sufficient to reduce the vocabulary size from 11 to 1, i.e., word "NINE".

For a larger size vocabulary, we need to design a proper method of merging diphthong labels with all the other labels generated from three critical segments, to maintain a meaningful phonetic code sequence for that path. For example, if both words "WHITE" and "RIGHT" are in the vocabulary, we need to keep the phonetic labels recorded from the critical segments preceding diphthong segment /AY/, and to see if these labels contain /W/ or /R/.

The performance of the diphthong recognition algorithm is discussed in the next chapter. By using the critical segment labeling algorithm and the diphthong recognition algorithm, all the non-nasal portions on a perceptual path can be automatically labeled. Thus far, we can represent an incoming utterance in a mixed form where both phonetic classes and phonetic labels are used. For all those non-SGS segments we have phonetic classes represented in the terms of their class names and class attributes such as lengths and average amplitudes. For those SGS segments (non-nasal portion) we have a sequence of phonetic labels. Collectively, we call them phonetic codes.

5.6 PHONETIC DICTIONARY

In SWIS, the phonetic dictionary is defined in terms of phonetic codes described in the previous section. The forms of phonetic codes depend on the phonetic classes with which they are associated. Generally speaking, there are only two classes of phonetic codes used in the dictionary.

The first class of phonetic codes is introduced to represent the non-SGS segments. For example, an /S/ segment in word "SEVEN" might be represented as "[*](>120)". This class of phonetic code has two fields, namely, bracket field and parentheses field. In a bracket field, a symbol represents the name of that phonetic class ("*" represents an unknown class, see section 4.2.) In a parentheses field an inequality imposes constraints on the length of that class. In the above example, the inequality >120 limits the phonetic code to representing only those unknown segments of 120 frames or longer.

The second class of phonetic codes consists of a sequence of phonetic labels included in a pair of brackets. Each phonetic label is listed in Table 5.1.

TABLE 5.1 Phonetic Labels Used in SWIS Phonetic Dictionary

Labels	Word Examples
IY	BEAT
EY	BAY
AY	BUY
I	BID
e	BED
@	BAT
a	BOMB
^	BUS
U	BOOK
u	BOOT
)	BOUGHT

The following notations are used in the construction of the SWIS phonetic dictionary. They are introduced for the purpose of illustrating the dictionary in a form that will help the reader understand the role it plays in the recognition process. In addition, there are usually many entries for a given vocabulary word. These notations are designed to make the dictionary more compact.

[ ] Phonetic class boundary

[**X**] **X** can be either a sequence of phonetic labels listed in Table 5.1 or one of the following phonetic class names:

Nasal	Nasal segment
Unstressed GS	Unstressed Glottal-Source segment
SE-dip	Sonorant Energy dip segment
Silent	Silent segment
*	Unknown segment

[X](>n)	The length of segment X must be longer than n msec;
{a}	Phonetic label a may be absent;
a|b	match phonetic label a or b;
<a|b>	if a is absent then b must be present, or vice versa;
?	match any one of phonetic labels in Table 5.1

The following examples show how a sequence of phonetic codes is constructed using the above notations and what they represent.

Example 1. [w<U|)>][Nasal] includes the following two sequences:

[wU][Nasal] or
[W)][Nasal]

Example 2. [{u|)}a][Nasal] represents the following:

[ua][Nasal] or
[)a][Nasal] or
[a][Nasal]

Example 3. [{I}e{u|^}] represents the following:

[Ieu] or
[Ie^] or
[Ie] or
[eu] or
[e^] or
[e]

Example 4. [e][SE-dip][Unstressed GS] represents that the strongest glottal source segment is an /e/ segment which is followed by a sonorant-energy dip segment and it in turn is followed by an unstressed glottal source segment.

With the notations defined above we are now ready to formally define the SWIS phonetic dictionary. First, we would like to point out that the design of the dictionary is based mainly on the phonetic information extracted from the stressed syllable in a word. That is,

every syllable or every phoneme in a word is not always used in the dictionary entries for that word. In other words, a vocabulary word may be defined only in the phonetic codes associated with its stressed syllable. Because these phonetic code sequences are matched to a unique entry in the vocabulary DIGIT, it is sufficient for the SWIS word generator to make the final decision without knowing the phonetic information of the other syllables in the word. Second, all the entries in the dictionary are derived from the analysis of the training data, i.e., from 88 phonetic class sequences (generated from the SWIS front-end system) and 88 perceptual paths (generated by the SWIS Phonetic Decoder.)

The SWIS phonetic dictionary consists of two versions, namely, standard and advanced. In the standard version of the dictionary, each entry for a word is copied strictly from one or more phonetic code sequences generated by SWIS for the training tokens. With this version, we are able to recognize all the tokens in the training databases at a word level of 100% accuracy. Because the size of the final testing database was much larger than the one of the training database, exceptional phonetic code sequences were to be expected. In order to reduce the number of exceptional phonetic code sequences that might be generated by SWIS for new tokens, we made some "educated guesses" based on all the available perceptual paths and the behavior of SWIS in the previous testings. The advanced version of the dictionary was so designed to handle exceptional cases. For example, we knew that a perceptual path for word "ONE" might miss both target zones /w/ and /n/, but, if it passes the first target zone /U/ and then /^/, then it has

already distinguished itself from all the other paths. That is, we added an entry [?U^?] for the word "ONE" which would recognize an exceptional path for the word "ONE". The standard and advanced dictionaries are listed in Table 5.2 and 5.3, respectively.

Table 5.2 SWIS Standard Phonetic Dictionary

Word		Phonetic Spelling
ZERO	-->	[IR] [IR<w|U|)>]
ONE	-->	[<w|u>^] [Nasal] [w<)|U>] [Nasal] [{)|u}a] [Nasal]
TWO	-->	[*](<140) [{I}u]
THREE	-->	[Ri]
FOUR	-->	[U{)}{R}]
NINE	-->	[Nasal] [AY] [Nasal] [Nasal] [AY] [*] [Nasal] [AY] [*]
FIVE	-->	[AY]
SIX	-->	[*](>130) [<I|e>] [*] [Silent] [*]
SEVEN	-->	[*](>130) [e{^}] [SE-dip] [Unstressed GS] [*](>100) [^] [SE-dip] [Unstressed GS]
EIGHT	-->	[EY] [*] [Silent] [*]
TEN	-->	[*](<120) [e] [Nasal] [*](<120) [{I}e{u|^}] [Nasal] [*](<120) [e] [*]

Note that the entries for the word "NINE" are listed before those for the word "FIVE". The order of word entries is carefully selected in

order to reduce the number of unnecessary substitutions incurred in word generation (see next section).

Table 5.3 SWIS Advanced Phonetic Dictionary

Word		Phonetic Spelling
ONE	-->	[?U<^|)>?] [?u<)|^>?] [w<U|)|a|^>?] [Nasal]
SIX	-->	[*](>100) [<e|^|I>] [*] [Silent] [*]
SEVEN	-->	[*](>100) [<e|^|I>] [SE-dip] [*] [*](>100) [<e|^|I>] [*] [*](>100) [e{^|u}]
EIGHT	-->	[*](<80) [{e|^}I] [*] [Silent]
TWO	-->	[*](<100) [I{e}u]
TEN	-->	[*](<100) [<e|^|I>]
NINE	-->	[Nasal] [<a|^>{e}I] [<a|^>{e}I] [Nasal]
FOUR	-->	[^<U|)>?]
FIVE	-->	[a?I]

5.7 WORD GENERATION

In this section, we present a simple lexical accessing scheme for word generation in SWIS. As defined in the previous section, each vocabulary word in DIGIT is represented by one or more entries in either or both versions of the SWIS phonetic dictionary. Each such entry is called a schema. Because each schema matches one and only one vocabulary word, all we need do is design an algorithm to match a sequence of phonetic codes generated from any utterance with the proper schema in the dictionary.

First, let's review the procedures we have described so far regarding how an incoming utterance is processed in SWIS to produce a sequence of phonetic codes. We use as an example an utterance of the word "SEVEN". Initially, it is divided by the SWIS front-end system into five segments:

[*] [Strongest GS] [SE-dip] [Unstressed GS] [*]

Then, the durations of [*] and [SE-dip] segments are computed to produce the first-level phonetic code sequence as follows:

[*]230 [Strongest GS] [SE-dip]57 [Unstressed GS] [*]22

Finally, the strongest glottal-source (SGS) segment is used to produce a sensory path which in turn is transformed into a perceptual path, and then a single phonetic code /e/ is generated from the perceptual path. Thus, the final sequence of the phonetic codes for the utterance becomes

[*]230 [e] [SE-dip]57 [Unstressed GS] [*]22

and it is this sequence that the lexicon-matching algorithm tries to match with the schemata in the SWIS phonetic dictionary. Generally speaking, SWIS always tries to match an SGS segment of a schema with the same in the sequence of phonetic codes. If it does not match, SWIS will try the next schema, until all the schema have been exhausted in both versions of the dictionary. If the two SGS segments match, SWIS will then start to match the rest of the items in the schema with their counterparts in the sequence of phonetic codes. Using the above example and Table 5.2, the reader will find that the first SGS segment in a

schema matching [e] is one for the word "SIX" in the standard version. Based on our statement, at this time, SWIS will try to match the first item preceding [e], which is an unknown segment [*] of 130 msec. or longer. It will match its counterpart in the sequence because one in the sequence is an unknown segment of 230 msec. long. Thus far, all the items preceding the SGS segment in that schema have been matched by their counterparts in the sequence (actually, there is only one item in this case.) Then, SWIS tries to match the first item ([*] segment) following the SGS segment in the dictionary against its counterpart in the sequence ([SE-dip] segment) and fails. Once an unsuccessful match is made, SWIS will select the next schema and repeat the matching process described above. At this time, the first schema for the word "SEVEN" in the standard version of the dictionary is selected. The reader himself may continue this exercise and find that all the items in the lexicon will be matched with their counterparts in the sequences. Therefore, the word "SEVEN" will be generated to represent the meaning of the incoming utterance as a final output of SWIS.

Now, we present the lexical accessing algorithm. Assume that SPC represents the Sequence of Phonetic Codes generated from an incoming utterance, NAME represent the word associated with a current schema, and WORD is the final output of the algorithm.

LEXICON ACCESSING ALGORITHM

1. Execute the following steps first for the standard version and then for the advanced version;
2. i=1;
3. L(i) : ith schema in the current version of the dictionary;

4. Compare the SGS in L(i) with the SGS in SPC;

5. If they don't match, then goto Step 9;

6. Try to match all the items preceding SGS in L(i) with their counterparts in SPC; If they don't match, then goto Step 9;

7. Try to match all the items following SGS in L(i) with their counterparts in SPC; if they don't match, then goto Step 9;

8. NAME(L(i)) ==> WORD; goto Step 11;

9. If ith schema is the last schema in the current version of the dictionary, then goto Step 2;

10. i = i+1; goto Step 3;

11. Stop;

For any SPC generated from an incoming utterance, there are only three possible outcomes from the execution of the algorithm, namely, "recognized", "substituted" or "rejected". If an SPC is matched correctly with one of the schema for the underlying word, then the utterance associated with that SPC is said to be **recognized**. If an SPC is matched with one of the schema for the other words, then the underlying utterance is said to be **substituted** by an incorrect word. If no entry in either version of the dictionary matches an SPC, then it is said to be **rejected** by SWIS. Note that SWIS will terminate the algorithm above at the first match. The overall performance of SWIS at word level is evaluated based on the definitions for the rates of recognition, substitution and rejection.

6. EXPERIMENTAL EVALUATION OF SWIS PERFORMANCE

In this chapter, we describe the results of the experimental evaluation of SWIS performance. The evaluation was conducted through testing SWIS with the database TEST2 consisting of 231 tokens from 23 new talkers (11 males and 10 females, see section 1.5.2 for the details.) The performance of SWIS on the testing database is defined in terms of recognition accuracies at three different levels, namely, phonetic class level, phoneme level, and word level, respectively. We will present the test results at each of the three levels in the next three sections. In the last section of this chapter, we will provide a subjective assessment on the overall performance of SWIS based on the results of the test.

6.1 CLASSIFICATION ACCURACY

As we have used classification procedures in the initial testing for SWIS (see Section 4.2.4), we compute the classification accuracy through detailed error analysis. The three types of error are defined to measure the performance of SWIS classification algorithms, namely, "miss", "insertion", and "substitution". Generally speaking, a miss-type error occurs at the region of an utterance where SWIS should have found an appropriate phonetic segment but missed; it is called an insertion-type error when SWIS locates more phonetic segments for an utterance than defined for the word corresponding to the utterance; and a substitution-type error means that SWIS locates a phonetic segment at the correct region but gives the wrong label for the segment. A few examples will clarify our discussion. Consider the word "ONE";

according to our lexical representation at phonetic class level (see Table 4.1), there should be a nasal-final segment in the utterances for the word. If SWIS fails to locate a nasal-final segment, one "miss" error will be counted under word "ONE". This can be easily explained in the following two lists of phonetic class sequences for the word "ONE".

```
SWIS definition   :   [*] [GS] [Nasal] [*]
SWIS actual output:   [*] [GS]  [*]
                             ||
                             vv
                   (nasal segment is missing)
```

An example of an insertion error is given below where SWIS generates more silent segments ([S]) and unknown segments ([*]) for an utterance of "EIGHT" than are defined for the word.

```
SWIS definition   :   [*][GS][*][S][*]
SWIS actual output:   [*][GS][*][S][*][S][*]
                                   --vv--
                                     ||
                      (two extra segments inserted)
```

A substitution-type error can easily be calculated based on the comparisons of both the predefined phonetic class sequences for a word utterance and the actually generated phonetic class sequences for that word. If there is any mismatch between the two sequences a substitution error will be counted. The following lists show a substitution error for an utterance of "SEVEN".

```
SWIS definition   :   [*][GS][SE-dip][Unstressed GS][Nasal[*]
SWIS actual output:   [*][GS]  [*]   [Unstressed GS]  [*]
                              ||                      ||
                              vv                      vv
                    (substitution error)        (miss error)
```

Note that there is also a miss-type error in the actual output sequence of phonetic classes in addition to substitution error. Clearly, all the three types of errors can occur in an output sequence of phonetic classes. After we compiled all the error counts using the procedures described above, we found that there is little correlation between the types and the number of errors and the talkers. Rather, the resulting errors depend on the vocabulary words. Table 6.1 lists the distributions of classification errors across both the vocabulary words and talkers' sex. The data listed in Table 6.1 is based on a testing trial with the final test database TEST2 where there are the 1012 phonetic classes from 23 new talkers (10 males and 11 females). Note that SWIS classification algorithms perform on female utterances just as well as on male utterances. This is encouraging because it seems that the classification algorithms developed in this project are insensitive to the talker's sex. The most miss-type errors are associated with

Table 6.1 DISTRIBUTIONS OF CLASSIFICATION ERRORS

WORD	MALE			FEMALE		
	Insert.	Miss	Sub.	Insert.	Miss	Sub.
ZERO	3	0	0	2	0	0
ONE	0	3	1	2	3	0
TWO	2	0	0	2	0	0
THREE	4	0	0	4	0	0
FOUR	0	0	0	0	0	0
FIVE	2	0	0	0	0	0
SIX	0	1	0	0	1	0
SEVEN	0	8	6	2	9	6
EIGHT	3	2	0	6	1	2
NINE	0	11	0	0	11	0
TEN	2	5	0	2	5	0
Total	16	30	7	20	30	8

words "SEVEN", "NINE", and "TEN" because SWIS often fails to find a nasal segment in the utterances for these three words. This is anticipated since the cues used for the nasal identifications are not very reliable. In fact, we have no knowledge that there exists any nasal identification algorithm used in speaker-independent speech recognition with a sufficiently high accuracy (>95%).

Note that the absolute number of the classification errors are counted to generate Table 6.1 Another way of looking at these errors is to calculate the rates of the errors of each type based on the total number of phonetic segments from all the utterances in the testing database. The resulting error rates of each type are plotted in Figure 6.1. The sum of all the three error rates is less than 13%, which is comparable to the previous 9% of the total error rates when the earlier version of the classification algorithms were tested against a substantially smaller database (1 male and 1 female.) Although the size of the final test database is ten times larger than that of our initial test database, the total error rates are in the same order of magnitude. However, this does not mean that the total error rates will still be below 15% when SWIS is tested using a larger database, perhaps, consisting of 200 or more talkers. In addition, the vocabulary used in SWIS is not phonetically balanced, in the sense that those hard-to-recognize phonetic classes are NOT uniformly distributed among the vocabulary words. For example, there is only one word containing a "SONORANT-ENERGY DIP" segment (word "SEVEN").

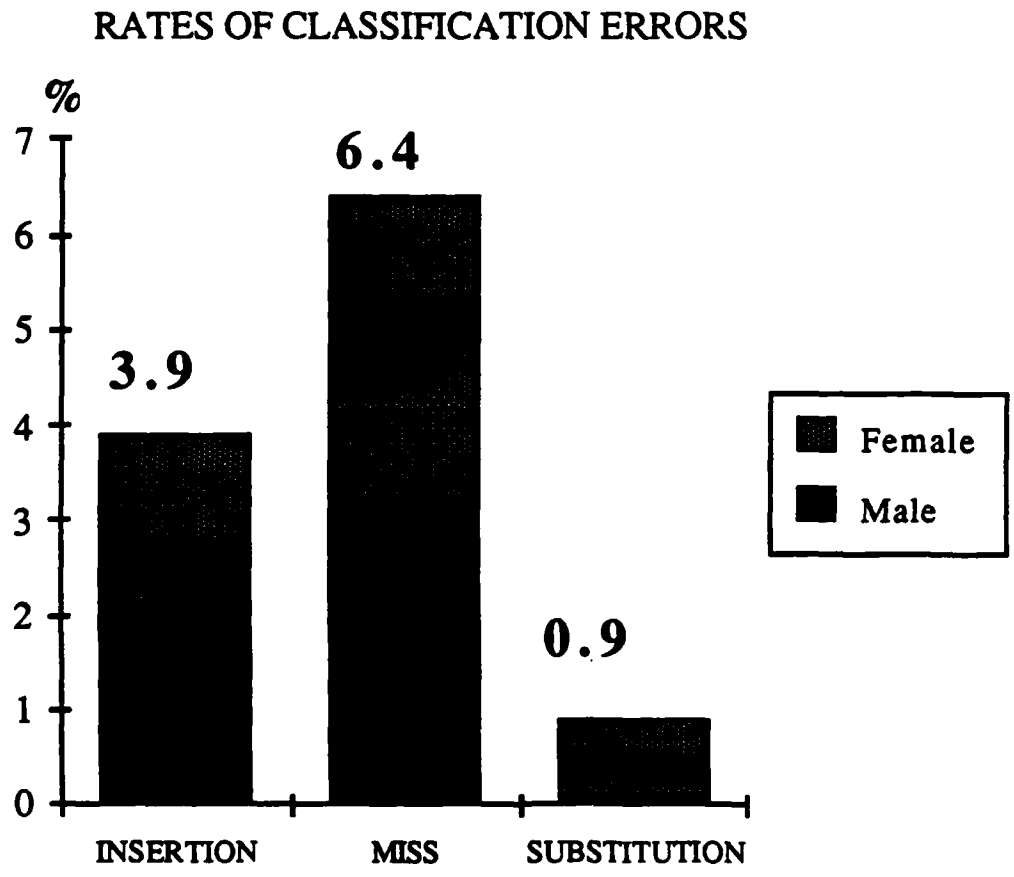


Figure 6.1 Error rates of the SWIS broad classifiers are ploted in each error category -- insertion error, miss-our error and substitution error. The three error rates in percent are based on the testing results from the database TEST2 which has the total 924 broad phonetic classes (44 broad phonetic classes in DIGIT and 23 new talkers recorded in the database TEST2.)

6.2 ACCURACY OF PHONEME RECOGNITION

One of the most important goals chosen for the SWIS project was to recognize the stressed syllables in DIGIT, more specifically, to recognize the vowels and semivowels in the stressed syllables. Therefore, the performance at phoneme level is considered more important than that at word level. Generally speaking, the performance of any phonetic recognizer is measured by the accuracy of phoneme recognition. In order to perform this measurement, a reference phonetic transcription for **EVERY** utterance, in whatever testing database is used, is needed. It is needed to compare with the actual phonetic transcriptions generated by a phonetic recongizer assuming that they are represented in the same phonetic alphabet, — for example, ARPAbet.

There are two problems with this method. First, it is very difficult to automatically generate reliable phonetic transcriptions from natural utterances. Of course, if we had such a system we would not need to develop a phonetic recognizer in the first place. So, experts must be called in to perform such tasks, and this directly causes the second problem. It is very expensive and time-consuming for human experts to produce phonetic transcriptions because (1) two or more well-trained phoneticians (to check for any human errors) are needed to do the job, and (2) they must listen to and write a phonetic transcription of every utterance in the testing database used. This latter approach is not practical for most small projects (such as the SWIS project) due to the limited resources. Even if a project has a budget for employing human experts, there will still be the problem of dealing with dialect difference. For example, given an utterance of

"ONE", if a phonetic recognizer generates a phonetic transcription /W+EH+N/, then how can we determine if /EH/ is the right phonetic label for the vowel in the word. We can not simply decide that the label /EH/ for the vowel segment is wrong because word "ONE" is transcribed as /W+AH+N/ in most English dictionaries. The reason is that these two phonemes /EH/ and /AH/ in many syllabic contexts sound so much alike. It is quite possible to have an equivalent result if we conduct an experiment where 100 phoneticians are called in to listen to those so-called "minimum pairs" (e.g., "ONE" vs. "WHEN") and then are asked to make independent judgment on what labels should be used in the transcription.

Because of the complexity and the required resources involved in the above method, we decided to adopt a simplified method of measuring the recognition rate and the insertion rate subjectively based on the phonetic codes generated by SWIS. The author carried out the measurement as follows. Given a sequence of phonetic labels generated by SWIS for a particular SGS (Strongest Glottal-Source) segment, we know the broad phonetic transcription for the stressed syllable corresponding to the segment. Then, we manually check for a phonetic label at a proper position in the sequence which matches the known phonetic transcription. If there is one, then we consider that particular phoneme represented by the label **recognized**. For instance, assuming that SWIS generates a sequence of phonetic labels /IH+R+UH/ for an utterance of "ZERO", we know that the first phoneme in the stressed syllable is /Z/ and the second is /IH/. Because the fricative /Z/ is not part of the SGS segment, the first phoneme for the SGS segment

should be /IH/ or /IH/-like. In this case, the first label is indeed /IH/ so we consider that this particular occurrence of /IH/ has been recognized. In this way, we can count how many occurrences of /IH/ have been recognized, by examining all the phonetic code sequences corresponding to the utterances which contain /IH/, i.e., examining all the utterances of "ZERO" and "SIX". The resulting recognition accuracies for /IH/ are 50% and 79% for male utterances and female utterances, respectively. This is shown in upper chart of Figure 6.2.

A phoneme is said to be recognized only if the corresponding phonetic label appears at the "right" position in the output phonetic code sequence. For the same word "ONE", if SWIS generates a different sequence than the one above, for instance, /UH+R+IH+UW/, then phoneme /IH/ in the utterance will NOT be considered recognized (in spite of the appearance in the sequence) because its position is too far (2 positions off) from the region where /IH/ should occur. Further, an insertion error will be recorded under the phoneme /IH/. In other words, phonetic label /IH/ appears in the place where it should not appear (e.g. "inserted" by SWIS.) For all the utterances whose SGS segments do not contain /IH/, we simply count all the occurrences of /IH/ in their phonetic labels as the number of insertion errors. For those utterances whose SGS segments do contain an /IH/ segment, we check the positions of each /IH/ occurrence to see if they are two positions off in the sequences. Only these two-position-off occurrences are counted as insertion errors. The insertion rate for a particular phoneme is then obtained by dividing the total number of insertions by the total number of the utterances not containing any occurrence of the phonemes in their

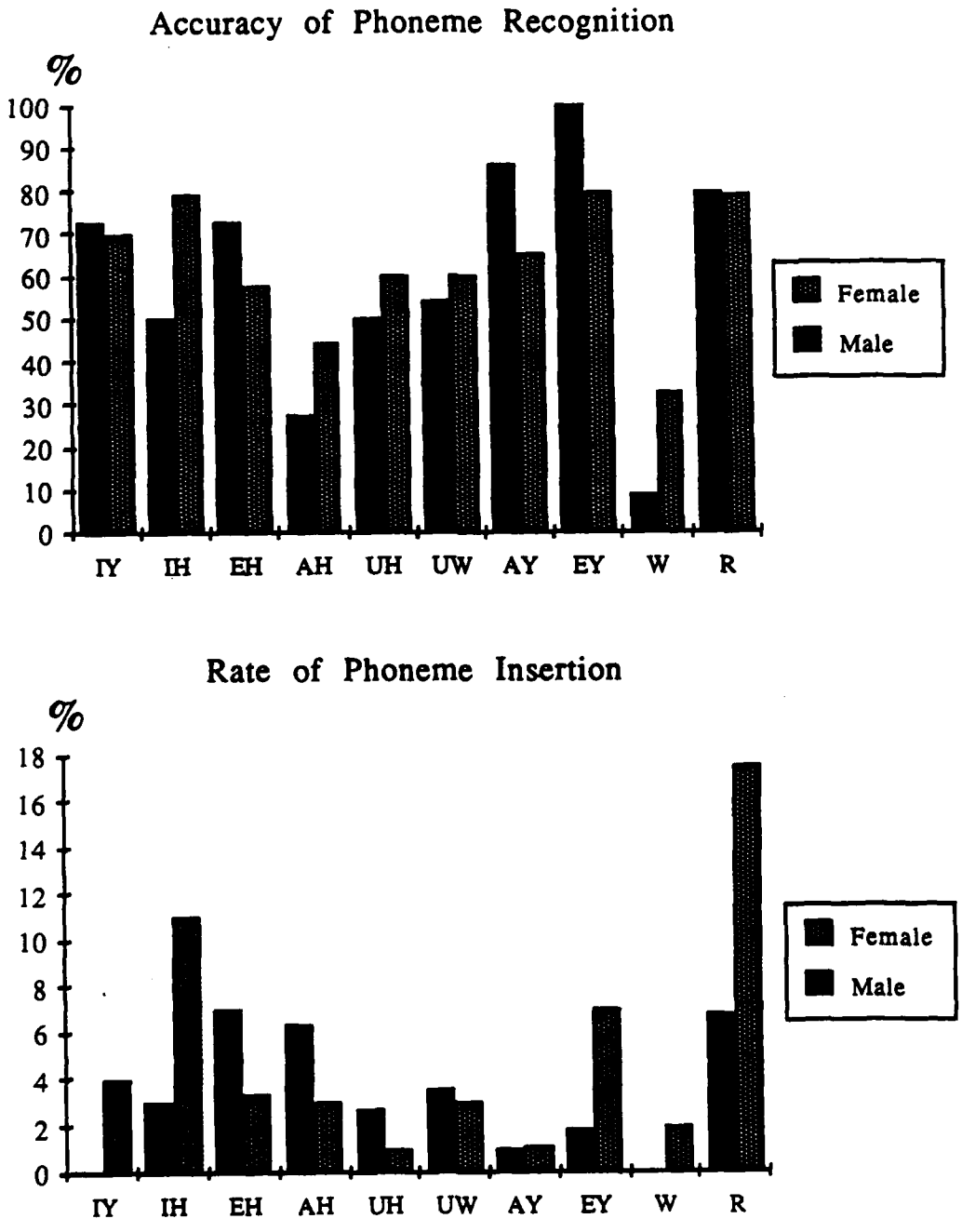


Figure 6.2 Recognition rates and Insertion rates plotted for each phoneme studied in SWIS are based on the testing results of all the utterances in the database TEST2.

SGS segments. As a result, the insertion rates for /IH/ are 3% and 11% for male utterances and female utterances respectively, based on our measurements on TEST2, as shown in the lower chart in Figure 6.2

Note that the recognition accuracies for all the phonemes except for /AH/ are greater than 50% (see Figure 6.2). The reason why the recognition accuracy for /AH/ is so low is because one syllabic context for /AH/ in DIGIT is /W+AH+N/ (word "ONE") and a great number of sensory paths for this word are not generated correctly by the SWIS front-end system. These invalid sensory paths are large in number, due to the difficulty of locating sensory formants for /W/ (SF1 and SF2, often shown as a single peak) and for /AH/ (nasalized by the following nasal consonant /N/).

Similarly, the high insertion rate for /R/ (18% for female utterances) is caused by the fact that SWIS often fails to locate the third sensory formant (SF3) so that it picks a location near the second sensory formant (SF2) for SF3. Because of the mathematical definition of the three-dimensional space APS, the perceptual path pointer will go below the so-called vowel-slab (Z'-coordinates are less than 0.6) as soon as SF2 and SF3 are close to each other. That is, the ratio of SF2/SF3 becomes very small. Because of our loosely defined target zone /R/, it is very easy for a perceptual path to enter the target zone /R/ because it is large (see section 5.3). In the original configuration for the target zone /R/ (89), the target zone /R/ is much smaller than the one used in SWIS. However, if that configuration is used, many perceptual paths would miss the target zone. Here, the tradeoff is between too many misses (if the target zone is too small) and too many

insertions (if the target zone is too large). Because we believe that the recognition of /R/ is more important (the /IH+R/ sequence is used to define word "ZERO") than having more phonetic codes containing an extra /R/ label, we decided to change the original configuration for the target zone /R/ based on the initial testing data in TRAIN and TEST1.

6.3 WORD RECOGNITION

As we stated in the previous chapters, the word recognition implementation is designed only to demonstrate that it is possible to construct a phonetic dictionary based on the phonetic codes (both phonetic labels for SGS segments and phonetic classes for other segments) and then to establish a lexical accessing scheme.

Word recognition accuracy is much more easily measured because there is no subjective judgment involved in deciding whether there exists a match ("recognition"), or mismatch ("substitution"), or no match ("rejection") at all between a SWIS-generated phonetic transcription and the phonetic dictionary constructed in SWIS project. Using the word generation procedure described in section 5.7, the final results obtained on the test database TEST2 (231 tokens from 11 unknown male talkers and 10 unknown female talkers) are shown in Table 6.2.

Table 6.2 WORD RECOGNITION ACCURACY

	MALE	FEMALE
RECOGNITION RATE	55.0%	52.3%
SUBSTITUTION RATE	11.6%	19.7%
REJECTION RATE	33.4%	28.0%

Because the total error rates (45% for male utterances and 47.7% for female utterances) are high, we decided to do a detailed analysis on all the perceptual paths whose phonetic code sequences are not recognized by SWIS. In other words, we wanted to know if these perceptual paths are valid or not, so that we could identify what parts of the SWIS recognition system caused the problems.

As a result of this perceptual path validation, we found that 66.7% of the total errors were caused by the SWIS front-end -- the classification and the sensory formant generation — in cases of the male utterances. Further, among the remaining 33.3% of errors, only the 11.3% were due to the logical errors in phonetic decoder (SWIS back-end), and the other 22% were correctable dialect errors, that is, they can be recovered by simply modifying the phonetic dictionary (without changing any SWIS software), such as adding a few more entries in the dictionary to accommodate the dialect differences. For the female utterances, the SWIS front-end is responsible for 80% of the total errors. Similarly, among the rest of the errors, only 9% are due to logical errors in the phonetic decoder and the other 11% are correctable dialect errors. One of the dialect errors encountered in the final testing is related to the phonetic definition for word "FOUR" (/F+UH+R/). For example, several phonetic code sequences generated from the word contain /AO/ (as in the word "WALL") rather than /UH/ so that they are rejected by SWIS. We could add to the phonetic dictionary the entry /F+AO+?/ for the word "FOUR" to match these sequences. However, such a quick fix will not work if the vocabulary contains the word "FALL" because it will increase the substitution rate.

Generally speaking, the term "logical errors" used in the previous section means that there are no visible errors on a perceptual path when it is plotted on the Phonetically-Encoded Auditory Perceptual map (P-map), see section 5.3 for details. In other words, that perceptual path has passed all the target zones it is supposed to pass (so, you can not blame the classification or the sensory formant generation for the problems) and the only reason for the incorrect phonetic code sequence generated from the path is that the phonetic parser chose the wrong location on the path.

The dialect errors are expected because the train-to-test ratio is less than 0.2 in this project (4 talkers in the training database and 21 new talkers in the testing database). Most of the other speaker-independent recognition systems reported in the literature used substantially larger training databases (the train-to-test ratio is often greater than 1) to develop the robust recognition algorithms. It is our belief that the dialect errors can be significantly reduced if we use a large training database consisting of 50 or more speakers.

Another fact contributing to the low recognition rate at the word level for the English digit vocabulary (DIGIT) is that most of the vocabulary words consist of single syllables. If the stressed syllables are missed by a recognition system then there is not much that can be done to recover from mistakes. Many agree that DIGIT is considered the vocabulary of second highest confusion, just below that of English letter vocabulary (26 English letters). Furthermore, the only difference between an /S/ and a /T/ reflected in the SWIS phonetic dictionary is that an unknown segment is called an /S/ if it precedes an

SGS segment and lasts longer than 100 msec., and is called a /T/ if it precedes an SGS segment and lasts no longer than 80 msec. Clearly, if we could truly identify the stressed syllables in DIGIT (note that SWIS tries only to recognize the vowels or semivowels in the stressed syllables in DIGIT) the overall recognition rate would be increased considerably. However, this will require generating a sensory path for all the segments corresponding to stops and fricatives (i.e., /S,T,Z/ for DIGIT). Unfortunately, so far we are not able to develop a successful algorithm to automatically generate sensory paths for non-glottal-source segments, although efforts are being made in several ongoing research projects at CID.

6.4 DISCUSSION

Based on the above described performance measurement, we summarize the results as follows. The overall classification errors are less than 12% (this is the worst case of performance figure because we assumed that we could not locate any nasal segments in unstressed syllables although we did not attempt it.) The overall phoneme recognition accuracy is about 61.5%. Although we are not completely satisfied with this figure, we consider the study a success, because there was only a 4.25% insertion error in our recognition rate. The recognition rate of 53.6% at the word level is far from acceptable for any stand-alone recognition system. Among the 46.4% of the recognition errors at the word level, 12% are due to the SWIS broad phonetic classifiers, 12.3% are due to the SWIS back-end system (the phonetic parser, the lexical accessing and the word generation), and the remaining 22% are caused by the SWIS sensory formant generator. This last figure means

that the ultimate accuracy obtainable from a formant tracker in the generation of sensory paths remains an open question. According to a recent report on the DARPA system developed at CMU (100), a new formant tracking algorithm based on the center gravity of DFT spectra promises an 80% accuracy for formant extraction in the vocalic segments, independent on syllabic contexts and speakers. However, we expect that an 85% accuracy for formant extraction will be the upper limit.

Obviously, the recognition of the vowels and semivowels contained in the stressed syllables, even in a small-size vocabulary like DIGIT is not sufficient to recognize every word in the vocabulary unless the phoneme recognition accuracy for these vowels and semivowels reaches 95% or more. It is questioable if one can ever build a phonetic recognizer with that kind of accuracy with very low insertion rate (less than 5%). Although one might argue that the small size of the training database used in this project and the limited time and manpower involved in the system development are responsible for the poor performance at the word level, we think that a system like SWIS could reach 90% word recognition accuracy at most.

Realistically, a SWIS-like recognition system can be used as a word hypothesizer in a larger recognition system, rather than working by itself, to generate a very small number of word candidate of good quality so that other system components can verify them, based on a different recognition philosophy such as dynamic programming wrapping (DPW) methods or vector code book techniques. For example, the phonetic

code sequences generated by SWIS reduce the number of word candidates from 11 to less than 2, on average, based on the results obtained in the final test.

On the other hand, if we can reliably identify consonants in the most stressed syllable (MSS) in a word utterance as we did for the vowels in the MSS, then we can significantly increase the recognition accuracy at word level. For example, with the vocabulary DIGIT, the total recognition accuracy will be increased if we could recognize stop consonants /S/ and /T/ in the vocabulary DIGIT.

7. IMPLEMENTATION

All the SWIS software was developed on a micro-VAX II under the VMS operating system. The software consists of two separately compiled packages, called FORMANT_TRACKING and FORMANT_PERCEPTION. Both packages are written in FORTRAN-77. There are about 6000 lines of source code in the packages. The software development began in October 1985 and concluded in August 1986. The entire effort is estimated have taken about 12-man-months.

All the datafiles shared by both packages are ILS-compatible, so that many intermediate results generated by SWIS can be accessed using ILS commands at the VMS operating system level. This file compatibility makes it possible to take advantage of using a variety of graphic display utilities (including graphic screen dump on most popular HP 7000 graphic plotter series) offered by ILS.

Once a waveform file has been stored on the disk, it will take FORMANT_TRACKING about two minutes (running SWIS alone on a micro-VAX II) to generate a sensory path for an utterance of 0.8 second. More than 90% of this elapsed time is used to compute 250 1024-point complex FFT in order to obtain short-term spectrum envelopes. Therefore, with a hardware implementation of complex FFF, the time spent on sensory path generation can be reduced to 12 seconds per 0.8 second speech, that is, 15 times real time for a non-optimal implementation on a non-parallel computing environment.

Once a sensory path is stored in an ILS data file (analysis file format), it takes FORMANT_PERCEPTION about three to five seconds to generate a phonetic code sequence for that path. Because we want the mapping between phonetic code sequences and the lexicon entries in the phonetic dictionary to be as transparent as possible, the word generation procedure is not coded. Instead, a set of rules are established, (see section 5.5,) to guide the mapping so that a novice user of SWIS can be trained in a few minutes to perform the word generation, given a phonetic code sequence output from SWIS. Table 7.1

Table 7.1 SAMPLE PHONETIC CODE SEQUENCES GENERATED FROM THE TEST DATABASE TEST2

WORD	Phonetic Code Sequences	sex
ZERO	[Silent][*]32[IRU][*]134[Silent]	m
ONE	[Silent][*]32[w^][*]35[Silent]	f
TWO	[Silent][*]102[u][*]25[Silent]	f
THREE	[Silent][*]22[Ri][*]19[Silent]	f
FOUR	[Silent][*]19[UR][*]80[Silent]	f
FIVE	[Silent][*]16[AY][*]70[Silent]	f
SIX	[Silent][*]137[I][*]25 [Silent][*]211[Silent]	f
SEVEN	[Silent][*]134[^][SE-dip] [unstressed GS][*]16[Silent]	m
EIGHT	[Silent][*]32[EY][*]22 [Silent][*]89[Silent]	m
NINE	[Silent][*]54[Nasal][AY] [Nasal][*]16[Silent]	m
TEN	[Silent][*]54[e][Nasal][*]12[Silent]	m

lists some actual phonetic code sequences output from the FORMANT_PERCEPTION package. All the sequences listed in Table 7.1 are successfully matched with a proper entry in the phonetic dictionary.

The 12 man-month effort was not evenly distributed between the two packages. In fact, the FORMANT_TRACKING package itself took almost eight months to complete. If we had anticipated the complexity encountered in the design of the phonetic decoder (PP), (see chapter 5,) we would have allocated more time to the software development for PP, as well as the design of the SWIS phonetic dictionary. Looking back at the time and the effort spent on the development of sensory formant generation (SPG) from the cost-performance point of view, we think that spending 75% of total development time for SPG is too much, considering that only about 65% of the sensory paths generated by SPG are valid. It is well worth mentioning why we spent such a great percentage of the total software development time on SPG. In the initial design of SPG, we took a quite different approach to formant extraction than the one used in the final version of SWIS implementation, that is, we used a digital-filtering-based spectrum processing technique. Because this technique was completely new to us and has never been explored in the context of formant extraction, we could only implement the technique in a trial and error fashion. In essence, this technique requires designing a set of digital filters and then applying them to short-term spectra in both linear- and log-scale (i.e., to treat the spectra as a time series) and hopefully to enhance the spectral prominences on the envelopes obtained from the filtered spectra.

Basically, there are four ways to try this technique, given a short-term FFT spectrum represented as M(i), where the value of M(i) is the frequency response of ith frequency component in log-magnitude (dB) at a particular frame assuming that the distance between each component is in linear-scale.

In the first method, we convert linear array M(i) to an equivalent log array $M_{L1}(i)$ through interpolation, and then apply a set of digital filters to $M_{L1}(i)$, and then extract sensory formants from the envelope computed from $M_{L1}(i)$. In the second method, we perform the filtering first on M(i) and then perform the linear-log conversion to obtain another time series, $M_{L2}(i)$, and finally extract sensory formants from the envelopes derived from $M_{L2}(i)$.

In the third and the fourth methods, we add a spectral-tilting function before and after the linear-log conversion, respectively. We tried each method with the short-term spectra selected from all the vowel nuclei in the utterances in TRAIN. Because we knew the approximate formant positions for each vowel, we could tell if the spectral peaks on the envelope computed using the above methods matched the expected formant positions. The problem was that after many tries with the different digital filters we were not able to design optimal bandpass filters to work for all the vowels. That is, there were either too many spectral peaks (the passband being too wide) or too few spectral peaks (the passband being too narrow). After two month's work, including the time to develop necessary support routines to do linear-log conversion and graphics display, we concluded that it was impossible in the available time to design a new formant extraction system based on

the technique, which could compete with the performance of more traditional LPC based formant extraction algorithms. Therefore, we changed direction and developed our current sensory formant generator which is built on several known LPC-spectrum based formant extraction algorithms.

There is no question that formant extraction remains one of most difficult problems in the design of any feature-based recognition system. We do not recommend that anyone try to improve the overall performance of SWIS or to develop a SWIS-like recognition system unless a better formant extraction algorithm has been developed and proved to be sufficiently reliable to generate formant contours from a speech database consisting of 300 or more speakers of many English dialects.

8. CONCLUSION

We have presented a new model of phoneme perception for the recognition of isolated words from an 11-word English digit vocabulary. A current version of this recognition system was tested using a 231 token database recorded from 21 new talkers (11 male and 10 female) and from which a 61.5% score for phoneme recognition with a 4.25% insertion rate and a 53% score for word recognition were obtained.

As far as the phoneme recognition accuracy is concerned, the reasonable success of this experimental system has answered the main question that it is designed to answer. That is, it is feasible to design an automatic speech recognition system based on the model proposed in the auditory-perceptual theory of phonetic recognition. Furthermore, several important factors that are crucial to the success of any future recognition system based on the theory, have been identified. Among them, a reliable sensory formant extraction algorithm is absolutely essential in order to achieve the high accuracy of phoneme recognition, and a phonetic parser which produces a very low number of insertion errors is needed to reduce the number of candidate words to 5% of the size of any active vocabulary.

In spite of the problem we experienced with formant extraction, we are optimistic about the outcome of future research in this area. We expect that the 85% of formant extraction accuracy can be achieved. However, it is generally agreed that the problem of recognizing non-vocalic consonants (i.e., stops and frictives) can not be sufficiently solved based on formant extraction. This means that we have to develop

new algorithms to automatically generate a sensory path corresponding to these non-vocalic sounds using either a signle auditory-perceptual space (APS) as we did in this project or multiple heterogeneous feature spaces, one for each phonetic class.

Because of the simplicity of the phonetic code representation used in the SWIS recognition system, great promise is shown that such a phoneme-oriented system based on this broad framework might provide a solution to speaker-independent isolated word recognition for a medium-sized vocabulary of 50 to 200 words. Furthermore, the research in this direction will help us to understand the fundamental relationships between the acoustic signals and phonetic elements which we and many other researchers believe will in turn help us to build a system recognizing natural speech.

9. ACKNOWLEDGEMENT

I would like to express my gratitude to Dr. James D. Miller for his time and effort in supervising this project. During my past four years at CID I have benefited from his expertise in acoustic-phonetics, his positive attitude to problems in scientific research, his optimistic thinking, and his commitment to always doing one's best. Besides thanking the readers for reading and offering valuable comments on the manuscript, I would also like to thank each one of them individually: to Dr. Maynard Engebretson for teaching me what speech processing is all about; to Dr. Marios Fourakis for his tutelage in linguistics; to John Hawks for collecting acoustic data used in the entire project; to Steve Sadoff for helping me in many programming aspects such as 2-dimensional contour generation; to Melissa Piasecki for formant labelling; to Kent Grant for sharing his thoughts with me. Most of the research described in this project was conducted at the CID research laboratories. To all of those at CID research laboratories who helped me along the way, I acknowledge their assistance with great appreciation.

Finally, I wish to thank my advisor, Professor Jerome Cox for his constant guidance and encouragement in my past five-year career of education at Washington University. It was his feeling about the importance of the applications of advanced computing technologies to basic scientific research that aroused my interest in the areas of speech recognition research.

10. APPENDICES

Appendix 10.1

The Auditory-Perceptual Theory of Phonetic Recognition *

James D. Miller
Central Institute for the Deaf
St. Louis, Missouri

10.1.1 Introduction

The purpose of this section is to present the current state of development of the auditory-perceptual theory of phonetic recognition.

Theories of speech perception treat the problem of how the acoustic waveform produced by a talker is transformed by the listener into linguistic units such as phones, phonemes, morphemes, words or other meaningful units. As previously noted by Miller (101), such theories can be described and compared in terms of a three-stage generic model.

In Stage 1 the acoustic waveform is transformed into auditory-sensory forms expressed in auditory-sensory dimensions. Although the analyses involved in Stage 1 can be described in a variety of ways, most commonly they are likened to short-term spectral analyses with parameters chosen to emulate, to a greater or lesser degree, the characteristics of the auditory system. Because of this, it is sometimes convenient to refer to "auditory spectra" or "sensory spectra." The notion of an auditory-sensory spectrum is, of course, very familiar and has been commonly used

* This section is taken from a grant proposal to the Air Force office of Scientific Research from the Central Institute for the Deaf (St.Louis, Missouri) dated January 26, 1986.

to interpret psychoacoustic data in terms of excitation patterns, and the works of Fletcher (102), Munson and Gardner (103), Plomp (104), Schroeder (105), Zwicker and Scharf (106), and Zwicker (107) serve as examples.

Stage 2 involves the transformation of auditory-sensory information into perceptually relevant dimensions. In motor theories it is here that sensory input is converted to articulatory terms such as abstract motor commands (Studdert-Kennedy et al, (108), Liberman, (109); Kozhevnikov and Chistovich, (110)). In feature theories, such as those of Fant (111); Stevens, and Blumstein, (112), or Pisoni and Sawusch (113), the auditory-sensory forms are converted to a perceptually relevant description in terms of auditory features. In the auditory-perceptual theory the auditory-sensory representation is converted to a higher-level, auditory-perceptual representation which serves an integrative-predictive function thought to be analogous to those visual-perceptual functions involved in perception of apparent motion, figure completion, and so on.

The final stage, Stage 3, involves the conversion of perceptual information into linguistic form. Here the perceptual information is converted to a form isomorphic with linguistic units. In motor theories, the motor commands or articulatory descriptions are converted to speech sounds such as phones or phonemes. In feature theories, the acoustic features are converted to phonetic features and then to phones or phonemes. In the auditory-perceptual theory, the dynamics of the perceptual

response in relation to previously established perceptual target zones causes the target zones to be activated and to issue neural symbols or category codes corresponding to the phones of a language.

Importantly, Stages 2 and 3 can be markedly influenced by "top-down" cognitive-perceptual processing wherein recent inputs including those from other senses and stored knowledge of language and events can have a significant influence. While Stage 1 may involve efferent neural activity from as high as the auditory cortex, such efferent activity is not considered to be the "top-down" processing referred to in discussions of speech perception as in those discussions a cognitive-perceptual aspect is implied.

It is within the broad framework described above that the auditory-perceptual theory will be considered. But before beginning that consideration, the reader is warned that a variety of new terms and concepts are to be introduced. These concepts and their interrelations constitute the auditory-perceptual theory, which is intended not only to have sufficient structure to have meaning but also to have sufficient flexibility so that it can be easily modified as required by the facts of speech perception. The concepts are meant to capture and include the explicit and implicit explanations of the past and to put them in terms that have the potential for an eventual quantitative evaluation. The reader is also warned that because the theory is undergoing development, its current state does not always

perfectly match its state in earlier abstracts, lectures, or manuscripts. Also it may be helpful for the reader to know the author's working assumptions concerning phonetic perception. It is assumed that when talker and speaker are native speakers of the same dialect, then the listener's response to carefully produced speech includes a series of internal category codes that are isomorphic with the phones of that dialect and that these are generated within the listener without top-down processing. However, whenever the process of phonetic perception -- up to the level of the phonetic category codes -- is degraded such as by filtering and noise or by mismatches between the dialects of the talker and speaker, then top-down processing becomes an important factor. Nevertheless, while trying to absorb the theory, it is recommended that the reader try to limit his evaluation to the special case of a listener with language and dialect matched to that of a careful talker and to the process of the conversion of an acoustic waveform to a phonetic string. Such a phonetic string is presumed to be a precursor to the perception of meaningful linguistic units such as words, for it is at a higher stage in the perceptual process, not dealt with by the auditory-perceptual theory, that the string of phones is converted into morphemes and meanings.

It is also useful to state that the auditory-perceptual theory is being developed in the context of the problems of segmentation, rate normalization, talker normalization, acoustic overlap of phonetic categories -- that is, the lack of acoustic

invariance — cue integration including bursts, transitions, and silence, coarticulation effects, and the mapping of auditory-perceptual dimensions on to articulatory dimensions.

10.1.2 Synopsis of the Auditory-Perceptual Theory of Phonetic Recognition:

The auditory-perceptual space. It is assumed that the sensory and perceptual responses to the input speech waveform can be located in a phonetically relevant auditory-perceptual space of only a few dimensions. The dimensions of this space are claimed to have characteristics similar to the variables $x = \log(P3/P2)$, $y = \log(P1/R)$, and $z = \log(P2/P1)$, where P1, P2, & P3 represent the frequency locations of the first three significant prominences in the short-term spectral envelope of the speech waveform as it is processed by the auditory sensory and perceptual systems. The variable R is a reference. As was reported earlier by Miller (114), R is the low-frequency reference and it is the relations of the positions of the spectral prominences to this reference that are the primary measures of the spectral patterns of speech. Thus, in this view, the values of the variables x, y, and z give the location of a spectral pattern in the auditory-perceptual space. When the reference and the prominences are thought to be at a sensory level, they are sometimes referred to as the sensory reference (SR) and the sensory formants (SF1, SF2, SF3). When the reference and the prominences are thought to be at the perceptual level, they are sometimes referred to as the perceptual reference (PR) and the perceptual formants (PR1, PR2, PR3). When the

distinction is unimportant or when the meaning is otherwise clear, the reference and the prominences are simply referred to as the reference (R) and the formants (F1, F2, F3). [Note that these values and concepts should not be confused with resonances of the vocal tract or true formants even though they sometimes correspond.] The locations of the phones of English appear to be organized in the space in a way that is simply and beautifully related to their phonetic and articulatory descriptions, and, in the case of the vowels, some of these relations have been described previously by Miller (91, 114-116). Of course, even though neither the particular dimensions given above nor their relations to phonetic or articulatory dimensions have been previously proposed, the general idea of such a space has been frequently stated or implied and the work of Peterson (117), Shepard (118), and Pols (119) provides examples.

Auditory-sensory analyses. It is hypothesized that the auditory system performs the equivalent of short-term spectral analyses based on the equivalent of a time-windowed waveform of 5-40 msec in duration. It is further hypothesized that these analyses produce the sensory equivalents of the amplitudes and frequencies of the tonal components in the input as suggested by the work of Scheffers (120) and at the same time these analyses separately produce the sensory equivalent of the continuous power spectrum of any significant aperiodic energy or other unresolved high-frequency components in the input waveform. This information is used to distinguish aperiodic, periodic, and mixed segments and

to establish the effective pitch (FØ) and "pitch strength" of the periodic and mixed segments. This same short-term spectral information undergoes further processing and, in this way, is used to generate auditory-spectral patterns that are variously referred to as sensory-excitation patterns, auditory-sensory spectra, or auditory-spectral envelopes.

Glottal-source spectra and burst-friction spectra. The analyses performed by the auditory-sensory system are hypothesized to produce two major classes of phonetically relevant auditory-sensory spectra. Members of one class have a prominence that can be associated with the first formant (F1). Such spectra are associated with sounds produced with a sound source at the glottis, whether periodic (voiced) or aperiodic (aspirated or whispered), and are referred to as glottal-source spectra (gs-spectra). Members of the other major class do not have a significant prominence in the region of F1. Such spectra are associated with friction bursts and sustained friction sounds and are produced with supraglottal sources. These spectra are referred to as burst-friction spectra (bf-spectra).

Sensory pointers. A glottal-source spectrum induces a sensory response in the auditory-perceptual space. This response can be thought of as a small object or a bundle of excitation and it is called a sensory pointer. The sensory pointer associated with a gs-spectrum is symbolized by GSSP for glottal-source sensory pointer. The location of this sensory pointer is $x = \log(SF3/SF2)$, $y = \log(SF1/SR)$, and $z = \log(SF2/SF1)$, where SR is

the sensory reference and SF1, SF2, & SF3 are the center frequencies of the first three sensory formants. A burst-friction spectrum induces a separate and distinct sensory response in the auditory-perceptual space. This response can also be thought of as a small object or bundle of excitation and it is called the burst-friction sensory pointer, symbolized by BFSP. For bf-spectra, the value of the absent first formant (SF1) is set equal to the current sensory reference, and the location of the burst-friction sensory pointer is $x = \log (SF3/SF2)$, $y = \log(SR/SR) = 0$, and $z = \log(SF2/SR)$. Thus, the burst-friction sensory pointer always lies in the xz-plane of the auditory-perceptual space.

Sensory paths. As the incoming speech is analyzed, the glottal-source sensory pointer (GSSP) materializes whenever a gs-spectrum is above the auditory threshold. As the values of SR, SF1, SF2, & SF3 change, the GSSP traces a sensory path through the auditory-perceptual space. The path of the GSSP is interrupted by silences and when the GSSP is replaced by the burst-friction sensory pointer (BFSP). One must imagine the GSSP moving through the space as the gs-spectrum changes shape and sometimes this movement is nearly continuous as in the case of the sentence, "Where were you a year ago?," where the only interruption would occur during the friction burst of /g/. In contrast, the burst-friction sensory pointer (BFSP) will usually appear and disappear as friction sounds are inserted in the speech stream. As bf-spectra are unstable, the BFSP may exhibit considerable jitter,

but it usually will not trace out a meaningful sensory path. In most cases where the BFSP replaces the GSSP, the BFSP will appear to "fill in" or "continue" the path of the GSSP as there are transitions in the path of GSSP to and away from the location of the BFSP. In the case of voiced fricatives, both sensory pointers are simultaneously present as one is associated with the gs-spectrum of the voiced part of the sound and the other is associated with the bf-spectrum of the friction part of the sound.

Spectrum goodness and spectrum loudness. Each sensory spectrum and, thus, each sensory pointer, is said to have a goodness and a loudness. The goodness is to be a measure of the "speech-likeness" of a sensory spectrum. The notion is that for each combination of SR, SF1, SF2, & SF3 an ideal speech-like spectrum can be defined. The cross-correlation between the ideal and the input spectrum will serve as a goodness index. Of course the correlation must have appropriate weightings of peaks and valleys [see discussion of Klatt (121) in Miller (101)]. The goodness index would be low for pure tones placed at the locations of the sensory formants, would be low for broad-band spectra with tiny bumps at the locations of the sensory formants, but would be high for carefully produced natural speech of high fidelity. A variety of schemes will produce adequate estimates of the loudness of a spectrum. For example, one could look-up the loudnesses of each of the formants and sum them with appropriate rules. A loudness index is defined to vary from 0.0 for below threshold

spectra to nearly 1.0 for for spectra that are comfortably loud as moderate loudness levels are known to produce near perfect intelligibility.

Thus, for each moment in time that a sensory pointer is above threshold it has a location in the auditory-perceptual space, a goodness or speech-likeness, and a loudness.

The perceptual response--the perceptual pointer. A perceptual response can be activated in the auditory-perceptual space by the sensory inputs (pointers). Like a sensory response, the perceptual response is also thought of as a tiny object or bundle of excitation in the auditory-perceptual space and it is called the perceptual pointer (PP). At each moment the perceptual pointer has an auditory state, a loudness, and a location. The auditory state of the perceptual pointer is related to the auditory states of sensory pointers. That is, it may be in a glottal-source state, gs-state; in a burst-friction state, bf-state; or in a dual, bfgs-state. The gs-state of the perceptual pointer may also be characterized as periodic (voiced) or aperiodic (aspirated or whispered) and as nasal or nonnasal. Similarly, the loudness of the perceptual pointer is related to the loudnesses of the sensory pointers. The location of the perceptual pointer is given by $x = \log(PF3/PF2)$, $y = \log(PF1/PR)$, and $z = \log(PF2/PF1)$, where PR is the perceptual reference and PF1, PF2, and PF3 are the perceptual formants. Since these perceptual variables are simply transformed values of the

corresponding sensory values, the location of the perceptual pointer is related to the locations of the sensory pointers.

The sensory-perceptual transformations for state, loudness, and location as integrative-predictive functions. It is assumed that the perceptual system automatically integrates sensory information over time in a manner that serves to optimize the perceptual representation of environmental events. That is, the perceptual system is thought to utilize sensory information in an integrative-predictive fashion.

In the case of the auditory state of the perceptual pointer, it is assumed that its auditory state matches those of the sensory pointers but that time is required for state switching. Such delays in state switching are assumed to range from 5 to 60 msec in duration and to depend on the loudnesses and sequencing of the sensory and perceptual responses surrounding the time of the switch of state. When both sensory pointers fall below threshold, the perceptual pointer retains its state for 100-200 msec, but then always returns to the most common state, the periodic, gs-state. Of course, it is hoped that the above concepts will prove useful in the explanation of the perception of voicing and aspiration.

In the case of the loudness of the perceptual pointer, it is assumed that the perceptual pointer almost instantaneously, that is within a few milliseconds, takes on the loudnesses of the sensory pointers. However, when the sensory pointers disappear, that is reach zero loudness, the loudness of the perceptual

pointer is thought to decay slowly over a period of 100-200 msec. In this way, the perceptual response is maintained during brief silences in the acoustic input.

In the case of the location of the perceptual pointer, the fundamental concept is that the sensory pointers attract the perceptual pointer and induce it to move through the auditory-perceptual space and trace out a perceptual path. Currently, this transformation is being modeled as follows. The sensory pointers are conceived as being attached by springs to the perceptual pointer.

The stiffness of a spring depends on the goodness index and the loudness index of its associated sensory pointer. In this way, near threshold spectra with little resemblance to speech would have almost no influence on the perceptual response while moderately loud speech-like spectra would have a strong influence on the perceptual response. Since a spring obeys Hooke's Law, the greater the separation between a sensory and perceptual pointer, then the greater will be the attractive force between them, and as the sensory pointer is not influenced by the spring, all of the force acts on the perceptual pointer. It is assumed that the perceptual pointer has mass and exhibits inertia. In addition, the auditory-perceptual space is thought of as being a viscous medium and the perceptual pointer thus encounters resistance. These concepts and variations of them can be readily cast in mathematical form and constitute the sensory-perceptual transformation for location.

When the sensory pointers disappear, the perceptual pointer begins to migrate to a "neutral point" that is appropriate to its auditory state. For example, if the perceptual pointer is in the burst-friction state when the sensory pointers disappear, it will begin to migrate to the burst-friction neutral point which is centered in the region of the auditory-perceptual space occupied by burst-friction spectra. Similarly, if the perceptual pointer is in the gs-state when the sensory pointers disappear, then it will migrate to the glottal-source neutral point which is centered in the region of the auditory-perceptual space occupied by gs-spectra. As stated earlier when the sensory pointers disappear, while the perceptual pointer maintains its state for 100-200 msec, it always eventually reverts to the gs-state. Therefore, in the case of a long silence, the perceptual pointer always returns to the glottal-source neutral point. As will be shown the concepts of the neutral points have the potential to play an important role in explaining the results of cue integration experiments and are essential for the appropriate segmentation of phones that are preceded or followed by long silences.

It is important to notice that sensory inputs including bursts, transitions, steady-states, and silences are all integrated into a single, unitary perceptual response by the hypothesized sensory-perceptual transformations. It is assumed that through experimental observations an appropriate choice of transformations and their parameters can be made and that considerable explanatory power may so be achieved. For example,

assume that the transformation for location is slightly underdamped. In this case, a sensory pointer could merely rapidly approach and veer away from a target location, and yet it could induce the perceptual pointer to overshoot and reach that desired location. This, as will be explained, is hypothesized in the theory for the case of nonsustainable consonants and may happen for a wide variety of cases in very rapid speech. Finally, it is noted that the idea of a perceptual system tracking and integrating sensory inputs is familiar as in the case of apparent motion in vision. In the case of speech, the notion has been frequently alluded to or implied and was explicitly stated in the work of Joos (122).

Auditory-perceptual events--"sounds." The perceptual response to sound, as described thus far, continuously moves through the auditory-perceptual space changing its loudness and auditory state as it goes. Such movement is meant to represent a continuously changing, unsegmented flow of auditory experience. It is asserted that such a continuous stream of experience can be automatically segmented into a series of discrete, auditory-perceptual events or "sounds" by an hypothetical segmentation mechanism. In the current theory, it is posited that the segmentation mechanism operates whenever the dynamic behavior of the perceptual pointer meets certain criteria. Previously, three candidate rules for such segmentation were proposed by Miller (123). These are: (a) an auditory-perceptual event occurs when the perceptual pointer undergoes a period of low velocity; (b) an

auditory-perceptual event occurs then the perceptual pointer undergoes sharp deceleration; and (c) an auditory-perceptual event occurs when the path of the perceptual pointer has high curvature. It was also suggested that perhaps all three candidates are correct when linked by or-statements, and that all three may need added time constraints such as a low velocity must be maintained for n-msec, or a path or a certain locus and curvature has to be traversed within certain time limits.

Perceptual target zones, neural symbols, category codes, the perceptual-phonetic transformation, and a reprise. The auditory-perceptual space is assumed to be divided into perceptual target zones. When activated, such target zones are capable of issuing distinct neural symbols or category codes. A perceptual target zone is activated when an auditory-perceptual event occurs within its boundaries. Otherwise said, when the perceptual pointer enters a target zone and its dynamic behavior operates the segmentation mechanism, then, and only then, is the target zone activated and a category code issued. While it is believed that target zones can be established for any class of sounds, it is the perceptual target zones that correspond to the phones of a language that are of interest here. It should be clear that the concept of the target zone is perceptual and not acoustic or sensory. For example, it is planned to conceptualize the target zones for stops as being physically unrealizable by letting the value of $y = \log(PF1/PR)$ be slightly negative. Thus, the spectrum and sensory pointer can only approach these targets and must do so

with appropriate dynamics such that the perceptual pointer will actually reach the target. Similarly target zones for sounds such as the consonantal /l/ in /lIp/ will be placed at points just beyond those normally reached by the sensory pointer. In other words, the targets for the nonsustainable speech sounds will be a placed outside of the normal range of the sensory pointer which will only be able go to "approach zones" in such a way as to induce the perceptual pointer to reach the more distant perceptual target zone. Of course, the target zones for the sustainable sounds such as the vowels may be entered by both the sensory and the perceptual pointers.

At present three slightly different types of perceptual target zones are being considered. The first and simplest type is one which defines a portion of the auditory-perceptual space which is uniquely associated with a single phone-like element. In this case, whenever an auditory-perceptual event occurs within that zone, a category code or neural symbol associated with that symbol is issued by the target zone. In the second type of perceptual target zone, the category code issued by the perceptual target zone is contingent on the auditory state of the perceptual pointer at the time of the auditory-perceptual event. For example, it may be that the perceptual target zones for the aspirated p, [ph]; the unaspirated p, [p]; and the voiced cognate b, [b] may show a substantial common area. In this case, the category code issued by the common area would depend on the auditory state of the perceptual pointer at the time of the event. A third type of

target zone might be particularly appropriate for glides and diphthongs. Here the target zone could be a unidirectional pipe that would be activated whenever the perceptual pointer enters one end of the pipe and exits the other within certain time limits, thus objectively defining quite literally a "perceptual glide." Preference for parsimony causes one to utilize the first and simplest type of perceptual target zone concept insofar as possible with glides and diphthongs defined as sequences of phone-like events. Data will surely force the issue and it is not unlikely that all three types of target zones will be required. In any case the basic nature of the proposed perceptual-phonetic transformation is clear. When the perceptual pointer performs a segmenting maneuver within a target zone, then the target zone issues a neural symbol or category code that corresponds to a phonetic element of the language.

The three stage process whereby the acoustic waveform of speech is converted into a string of category codes that correspond to a string of phonetic elements has now been presented. A brief reprise follows. At each moment the spectral envelope patterns of glottal-source sounds and burst-friction sounds are represented as sensory responses or sensory pointers in a phonetically relevant auditory-perceptual space. These sensory responses are converted to a unitary perceptual response by integrative-predictive transformations, sensory-perceptual transformations, that rely heavily on the histories, trajectories, and dynamics of the sensory responses. Finally, segmentation and

categorization mechanisms that depend on the dynamics of the perceptual pointer in relation to perceptual target zones result in a string of neural symbols or category codes that correspond to the speech sounds of a language.

In the remainder of this synopsis, the portions of the auditory-perceptual theory that deal with (a) the development of target zones, (b) the speech mode and levels of processing, and (c) the integration of top-down processing into the theory are presented. These provide background for, but are not directly relevant to, the current proposal.

Traces, tick marks, clouds of ticks, and the development of perceptual target zones. The concepts that are intended to aid in the understanding of the development of phonetic perception, cross-language differences, as well as selective adaptation experiments are now introduced.

Each auditory-perceptual event is said to leave a trace or a tick mark in the auditory-perceptual space. Perhaps these traces or tick marks have a quality that depends on the event that produced them or, perhaps they are simply defined by the coordinates of the events that produced them. In either case, when a cloud of ticks — that is, a region with a high density of tick marks surrounded by a region of lower density (perhaps of a given quality)—is formed, then, it is postulated, that the nervous system automatically places an envelope around that cloud and creates a perceptual target zone that is capable of being activated and issuing a neural symbol or category code.

Since traces or tick marks are assumed to fade or decay in time, it would usually be necessary to present a stimulus frequently and rapidly to create a cloud of ticks. Therefore, under most circumstances the target zones so created would be temporary and dissolve with time as the tick marks fade away. Other target zones, such as those for the phones of one's native language and dialect, are formed during infancy and childhood under certain circumstances, as yet unspecified, such that they are nearly permanent and difficult to modify. Perhaps this happens in a manner somewhat similar to the sensory imprinting suggested by Marler and Peters (124). In any case, the idea of nearly permanent and difficult to modify perceptual target zones corresponding to the phones of one's native language and dialect seem to be required by cross-language differences.

Since it is assumed that temporary target zones can be formed anywhere in the auditory-perceptual space by repeating a stimulus frequently in a short period of time, and if it is further assumed that such temporary target zones can subdivide larger existing ones, then it is tempting to use these concepts to explain the selective adaptation experiment. For as is known, category boundaries shift toward the adapting stimulus in such experiments by Cooper (125) and Diehl (126) as might be predicted if a uniform adapting stimulus resulted in creation of a small target zone around the cloud of tick marks produced by the rapidly presented adaptor. Of course, Diehl's (126) experiments and interpretation

speak against this view and his experiments do seem to require another explanation.

No matter the final explanation for the selective adaptation experiment, the concepts offered here provide a mechanism whereby perceptual events can be converted into learned categories by the development of perceptual target zones that are created by a system that is primarily driven by the stimuli presented to it.

Levels of processing -- the categorical level (speech mode) and other auditory-perceptual levels. Liberman and his coauthors (Liberman (109), and references cited therein) have repeatedly emphasized that speech-like stimuli can be processed by human listeners in different ways or modes. They refer to a speech mode as opposed to an auditory mode and they rally conclusive evidence in support of this kind of distinction.

Below it is suggested that auditory-perceptual information may be processed (a) at the categorical level, that is, in terms of neural symbols or category codes; (b) at the level of the auditory-perceptual event; or (c) at levels of the perceptual and sensory paths. The categorical level is intended to include Liberman's speech mode as well as the categorical mode for nonspeech material that can be similarly processed (Miller et al., (127)). The other levels discussed below appear to be subdivisions of his auditory mode.

In the auditory-perceptual theory it is posited that the information from the auditory-perceptual space can be forwarded in various forms to other perceptual-cognitive structures for

additional processing. For example, the information can be forwarded as neural symbols issued by target zones which are categorical in nature and place few demands on the listener's perceptual-memorial resources. If more detail is needed, the coordinates of the auditory-perceptual event can be used for additional processing. At an even more detailed level, a highly trained and attentively focused listener may utilize whole segments of the perceptual path, which place considerable demands on his/her perceptual-memorial resources. A fourth possibility is that, under special circumstances of training and attention, a listener may be able to utilize information from the sensory pointer directly.

Of course, the concepts introduced above are intended to be helpful in the explanation of the variety of results obtained in speech-perception experiments as the task demands and degree and kind of training are varied.

Top-down perceptual-auditory processing. The theory as just presented does not include the concept of "top-down processing." Indeed, it assumes that when listening to speech that is carefully produced by another native speaker of the same language and dialect no top-down processing is required for accurate phonetic perception. However, the importance of top-down processing in a great many listening situations cannot be denied, and the separation of the perceptual and sensory aspects of phonetic perception leaves ample opportunity for top-down processing to be integrated into the theory. For example, the listener's

expectations could be conceived of as adding forces that attract the perceptual pointer toward particular target zones. In this way, the perceptual pointer is driven not only by the sensory pointers and the other factors previously mentioned in the discussion of sensory-perceptual dynamics, but also by the listener's expectations as they are controlled by context, knowledge of the language, and so on. Another similar form of top-down processing could involve information from other senses resulting in attractive forces on the perceptual pointer. For example, mouth movements such as observed in lipreading could also result in the addition of forces that attract the perceptual-pointer to various target zones and thus influence phonetic perception. Finally, more complicated forms of top-down processing can be imagined. For example, the sizes and shapes of the target zones could be changed depending on the speech characteristics of the talker such as having a foreign accent, deaf speech, and so on. Of course many other kinds of top-down processing can be introduced as the output of the auditory-perceptual space undergoes additional processing such as that required for the identification of words and meanings.

10.1.3 Perceptual Target Zones

Data are being gathered in the laboratory and from the literature so that preliminary estimates of the sizes, shapes, and locations of the perceptual target zones for the phonetic elements of English can be made. Here it is the boundaries of enclosed volumes in the space defined by x=log(PF3/PF2), y=log(PF1/PR), and

z=log(PF2/PF1), where PR is the perceptual reference and PF1, PF2, and PF3 are the perceptual formants, that are of interest.

For the purpose of establishing the preliminary versions of the perceptual target zones we have reverted to the standard method of speech research wherein the investigator selects the appropriate segment in an a posteriori fashion, based on all available knowledge including the sound of the segment. Although this method is to be replaced by a completely automatic, "hands-off" procedure in the future, for now, the more usual method allows us to move forward and gather relevant information.

In placing points in the auditory-perceptual space, we use several novel concepts and tools that require explanation. The first is the concept of the reference. The concept of the reference is a key innovation and is crucial for our work. The idea is that the pattern of frequency locations of the prominences in a speech spectrum is described by the relations of the frequency values of these peaks to a reference frequency. The reference frequency is always defined, but is shifted slightly by the average spectrum of the current speaker and by significant pitch modulations. We seek algorithms that quickly adjust the reference so that the average logarithmic distance between the reference and the talker's average spectrum will remain constant across speakers of differing sex, size, pitch, and age. The current formula is based on the talker's average voice pitch and

on his pitch modulations between about 1.5 and 50 modulations per second. The sensory reference (SR) is calculated as

$$SR = 168(GMTFO/168)^{1/3} + FIL(MOD\ FO), \quad (1)$$

where GMTFO is the estimated geometric mean of the current talkers FO, and FIL(MOD FO) is the value of his/her modulation of FO after band-base filtering. For estimating points for target zones, the last term in the equation (Eq. 1) is usually ignored. Furthermore, when the pitch associated with the spectrum is not given, we seek other data of the talker's average pitch, or else we assume GMTFO is 133 Hz for men and 225 Hz for women. These correspond to references of 155 Hz for men, and 185 Hz for women.

For the assignment of numbers to the sensory formants we are guided by a set of rules that are undergoing evaluation. We distinguish glottal-source spectra and burst-friction spectra, primarily on the criterion of whether or not there is a significant first sensory formant. The significant peaks in a glottal-source spectrum are labelled as sensory formant one low (SF1L), sensory formant one high (SF1H), sensory formant two (SF2) and sensory formant three (SF3). These assignments are made in accordance to the locations of three bands (B1, B2, and B3) of log ratios of the frequency locations of spectral peaks to the frequency of the sensory reference. Band 1 extends from 0 to about 0.8; Band 2 extends from about 0.6 to 1.20; Band 3 extends from about 1.0 to 1.4. Our rules allow the first sensory formant

to be divided into two when either a band with a single peak is too broad, or when two peaks, neither of which is SF2, occur in Band 1. Under these conditions we speak of sensory formant 1 low (SF1L) and sensory formant 1 high (SF1H), and this is typical of nasal or nasalized spectra. Sensory formant 2 (SF2) is usually easily identified as being in Band 2. However, when no peak is found in Band 2, as may be the case in /w/, SF2 is set equal to SF1H or SF1, whichever is higher. SF3 is taken as the peak falling in Band 3. When no peak is present in Band 3, SF3 is arbitrarily set to $SR(10^{1.18})$, which is average value of SF3 for that reference value. The merging of formants usually is readily identified. For example, when the only peak in Band 2 is the same peak as the only one in Band 3, then SF2 and SF3 are taken as merged and equal, which is common in velar articulations. Similarly, when there is only one peak in B1 and it is the same as the only peak in B2, we say that SF1 and SF2 are merged and equal, as can occur with the vowels /ɑ/ and /ɔ/. Also, as mentioned earlier, when there is no peak in Band 2, then SF2 is set equal to SF1 or SF1L, whichever is higher, and SF2 is said to be merged with that formant. These relations and rules are illustrated by specific examples of glottal-source spectra below. In the case of burst-friction spectra, there is no peak in the first formant band. Two new bands are identified. Band 2 ranges from about 0.6 to 1.45 and Band 3 ranges from about 1.0 to 1.65. The lowest significant in Band 2, is labelled SF2, while the next highest

peak, that falls in Band 3 is labelled SF3. Examples of these will also be illustrated in the text that follows.

10.1.3.1 Simple Vowels: /iɪɛæʌɑʊɔu/.

Presently, the perceptual target zones for these vowels are the best established. In careful pronunciation the sensory and perceptual formants can be taken as identical. We have plotted data from several sources. These include means and individual measurements. Sources for mean data include: Peterson and Barney (128), male, female, child; Fairbanks and Grubb (129), male; and Klatt (89), male synthesis rules. Sources for individuals include: Peterson (129), 4-5 males, 3 females, and 3 children; Miller (130), one point per vowel for one male and one female in /bVb/ context; Lehiste (131), one point per vowel from one male (GEP); Miller and Chang (In preparation), 64 points for each vowel of the set /ɪɛʌ/; and Fourakis, Hawks and Miller (132), 2 males and 2 females for intoned vowels and 1 male and 1 female for vowels in CVC-sonorant context. These diverse data allow us to plot between 22 and 89 points per vowel.

It has been found by Miller (91) that the vowels fall in a narrow slab centered around the sum plane $(x + y + z) = 1.18$. The width of the slab is quite narrow, being of the order .12 log

units or 0.4 of an octave. A simple rotation of the axes places the vowel slab in the vertical where,

$$x' = .70711(x - y),$$

$$y' = .81622 - .4081(x + y), \text{ and}$$

$$z' = .5772(x + y + z).$$

Often we shall show front and side views of the vowel data plotted in "slab coordinates," that is, x', y', and z'. Please note that the location of spectra in the auditory-perceptual space, whether in xyz- or x'y'z'-coordinates, is a way of describing formant patterns or as Fant (111, 133-134) calls them F-patterns. Each point in the space represents a different F-pattern.

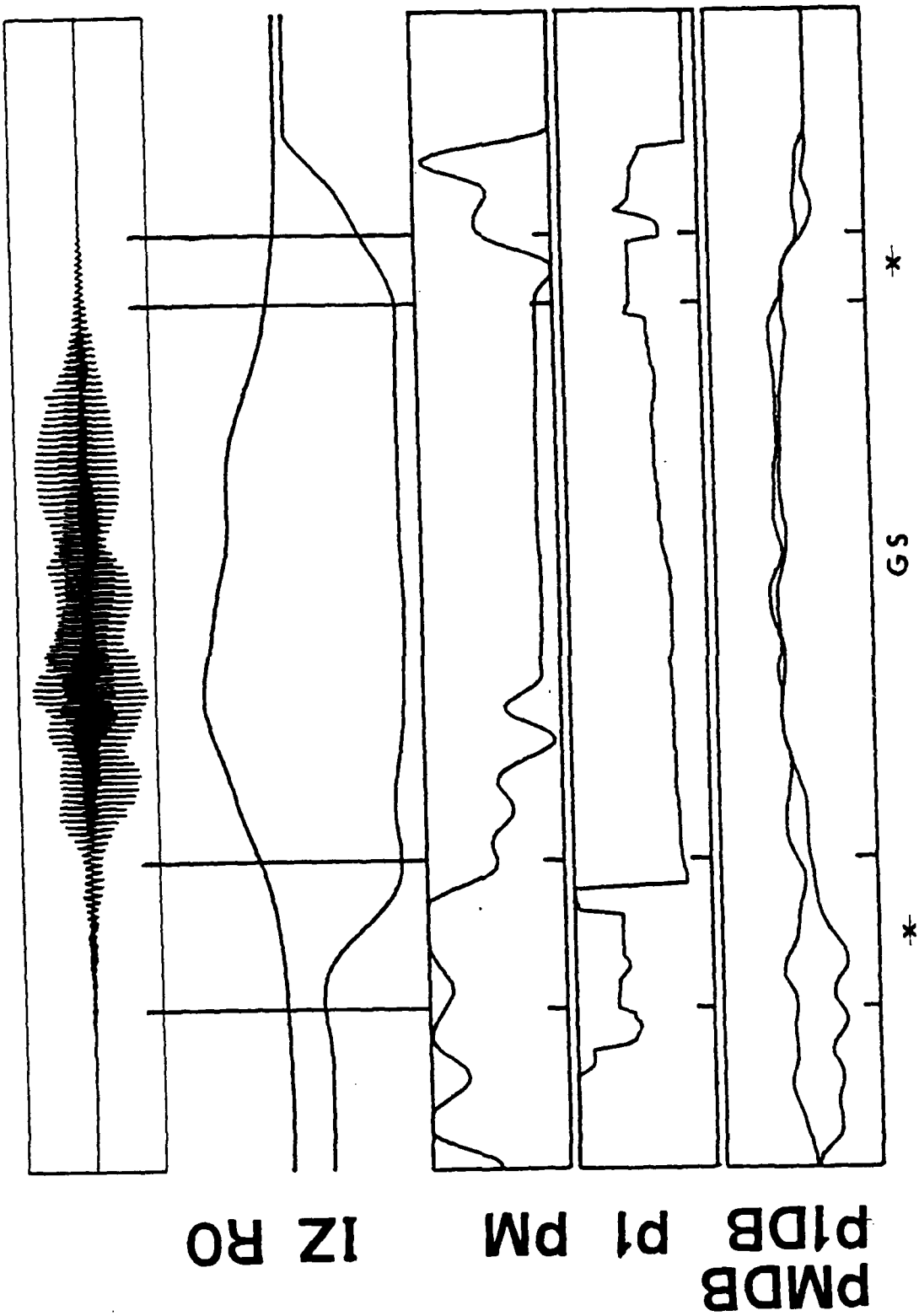
APPENDIX 10.2

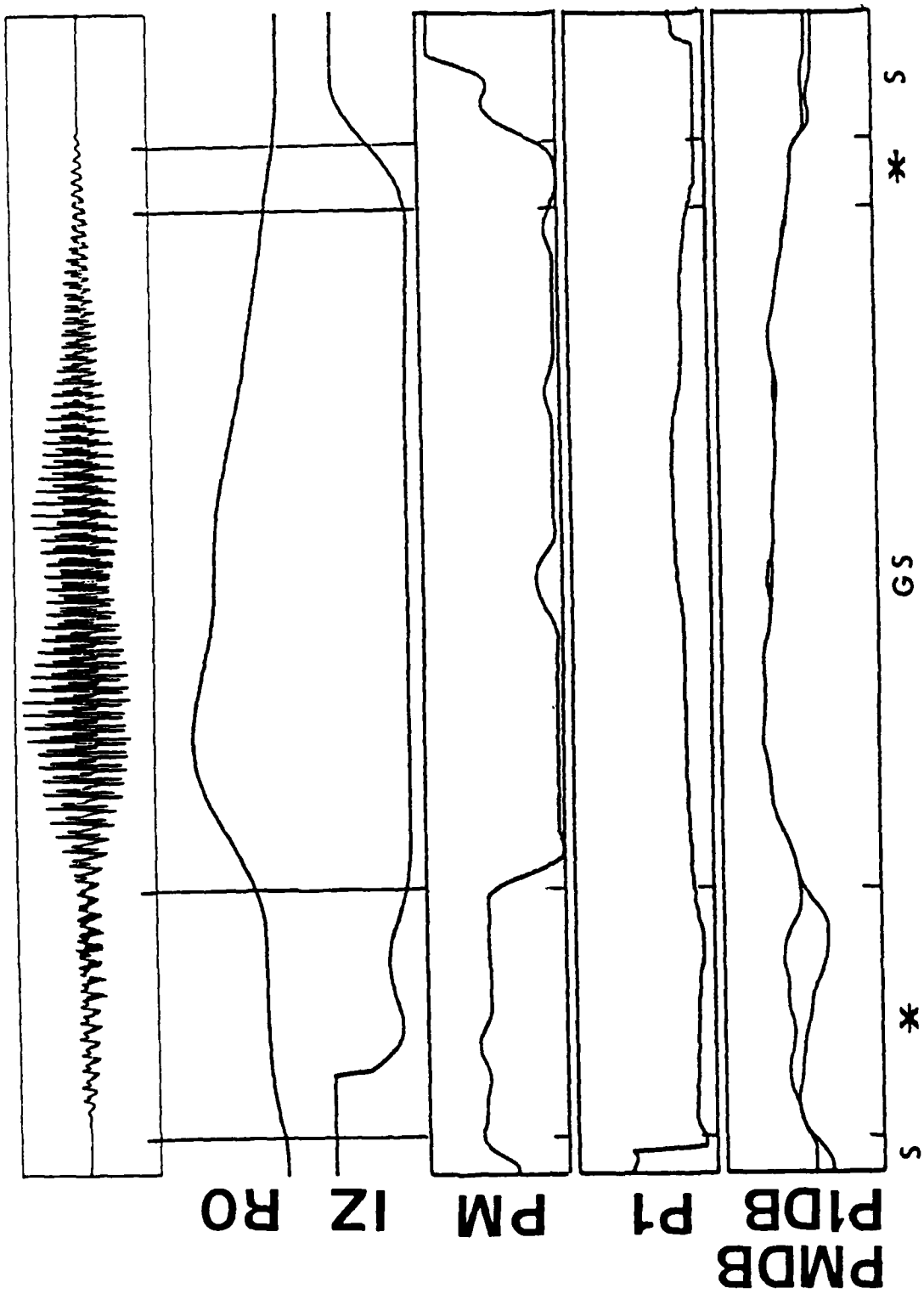
Classification Plots for Every Vocabulary Word

The duration of each of the following plots is 224 frames (3:226), i.e., 716.8 msec. The following labels are used to denote the broad phonetic classes:

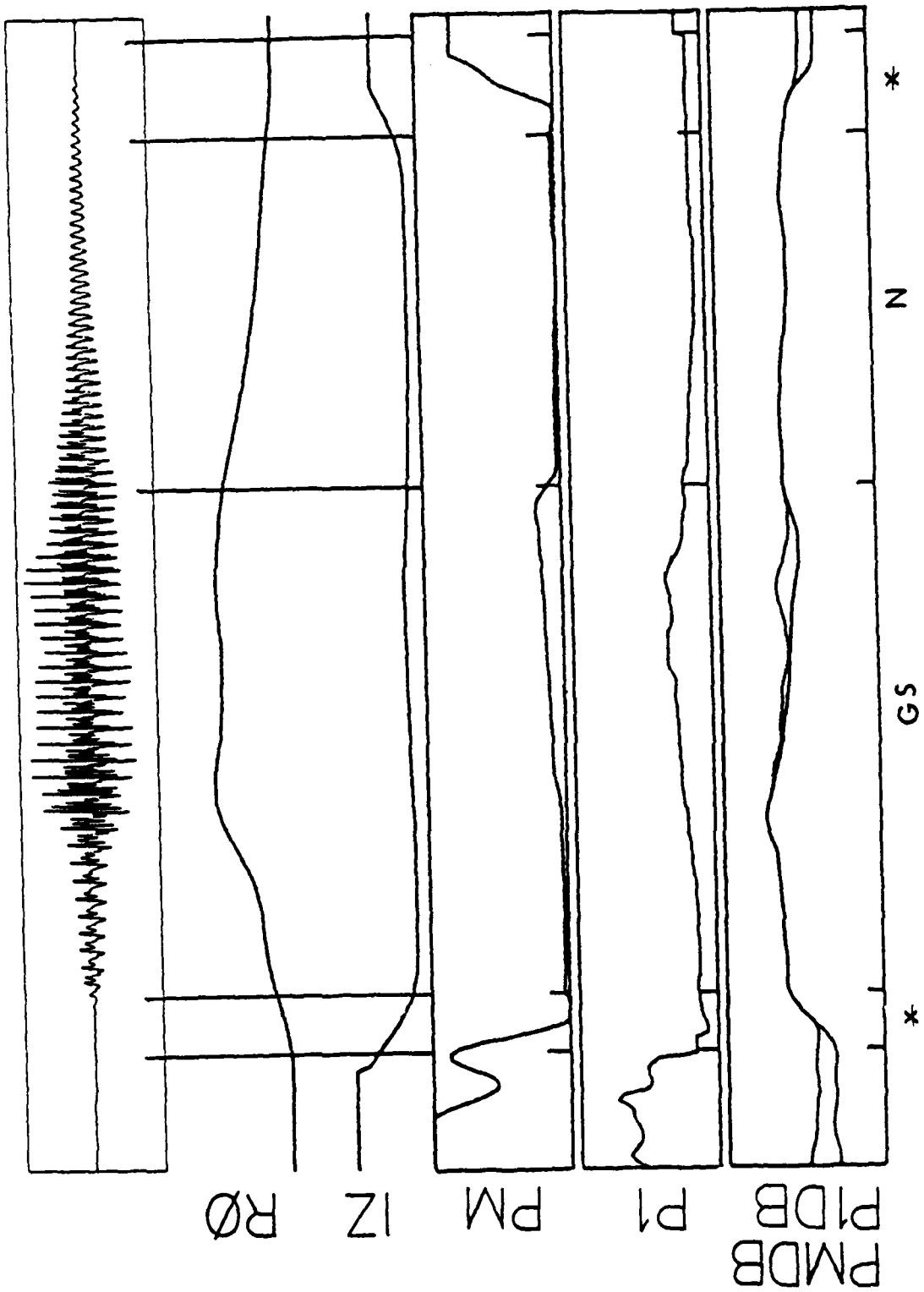
S	:	Silent
GS	:	Glottal-Source
D	:	Sonorant-Energy Dip
N	:	Nasal
*	:	Unknown (wildcarding)

Except for the first plot, the initial and the final silent segments are not marked. The vertical axis for PlDB and PMDB is in dB, ranging from 0 to 100. The vertical axis for Pl is in Hz., ranging from 0 to 2000. The vertical axis for PM is in Hz., ranging from 0 to 8000. There is one plot for each talker in the training database TRAIN (the first utterance is used for each of these two talkers.)

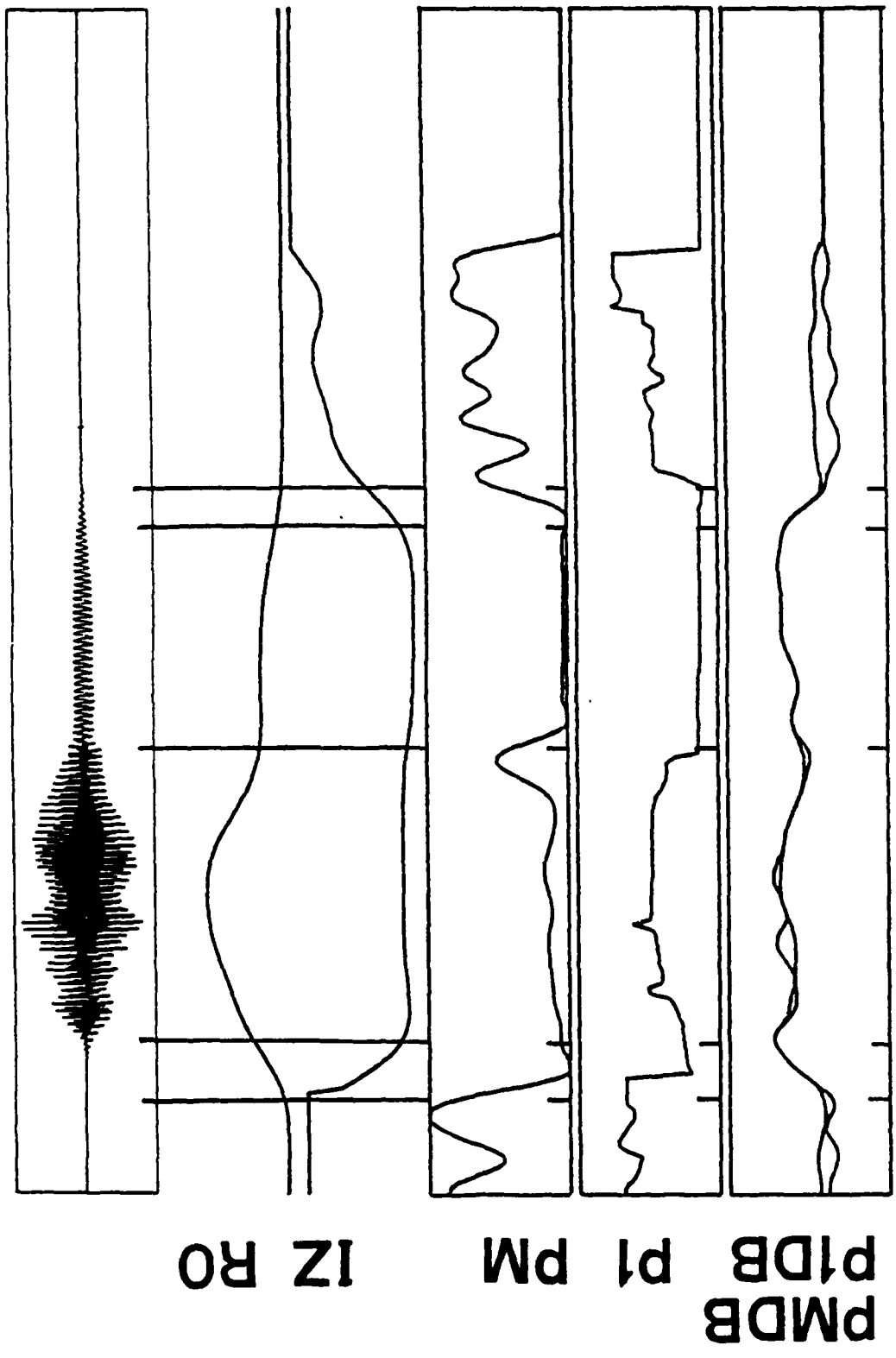

pMDB
p1DB
p1
pM
1Z
RO
GS
ZERO
(female: LT, wd7022.)



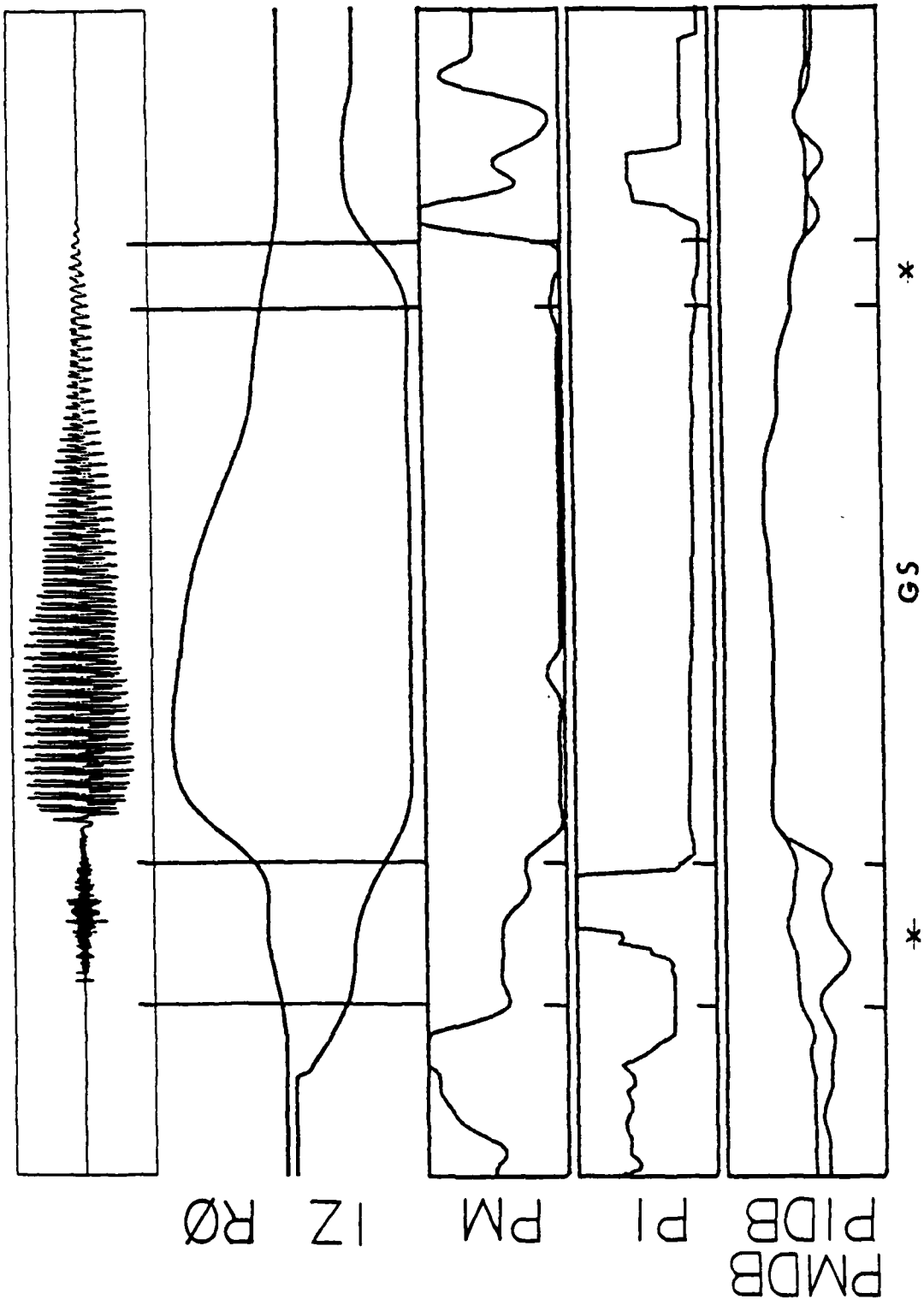
ZERO (male: MO, wd7000.)

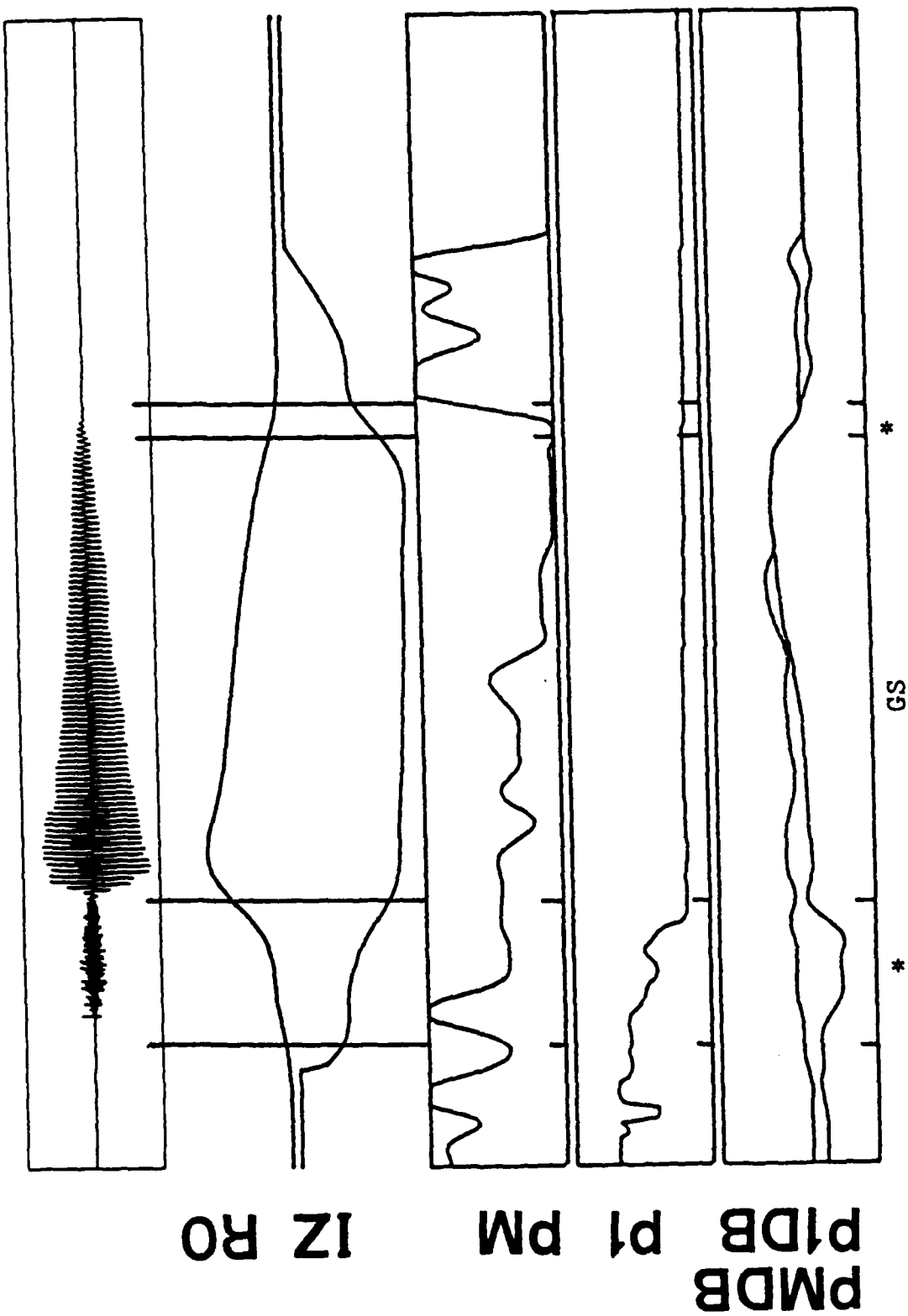


ONE (male: MO, wd7001.)

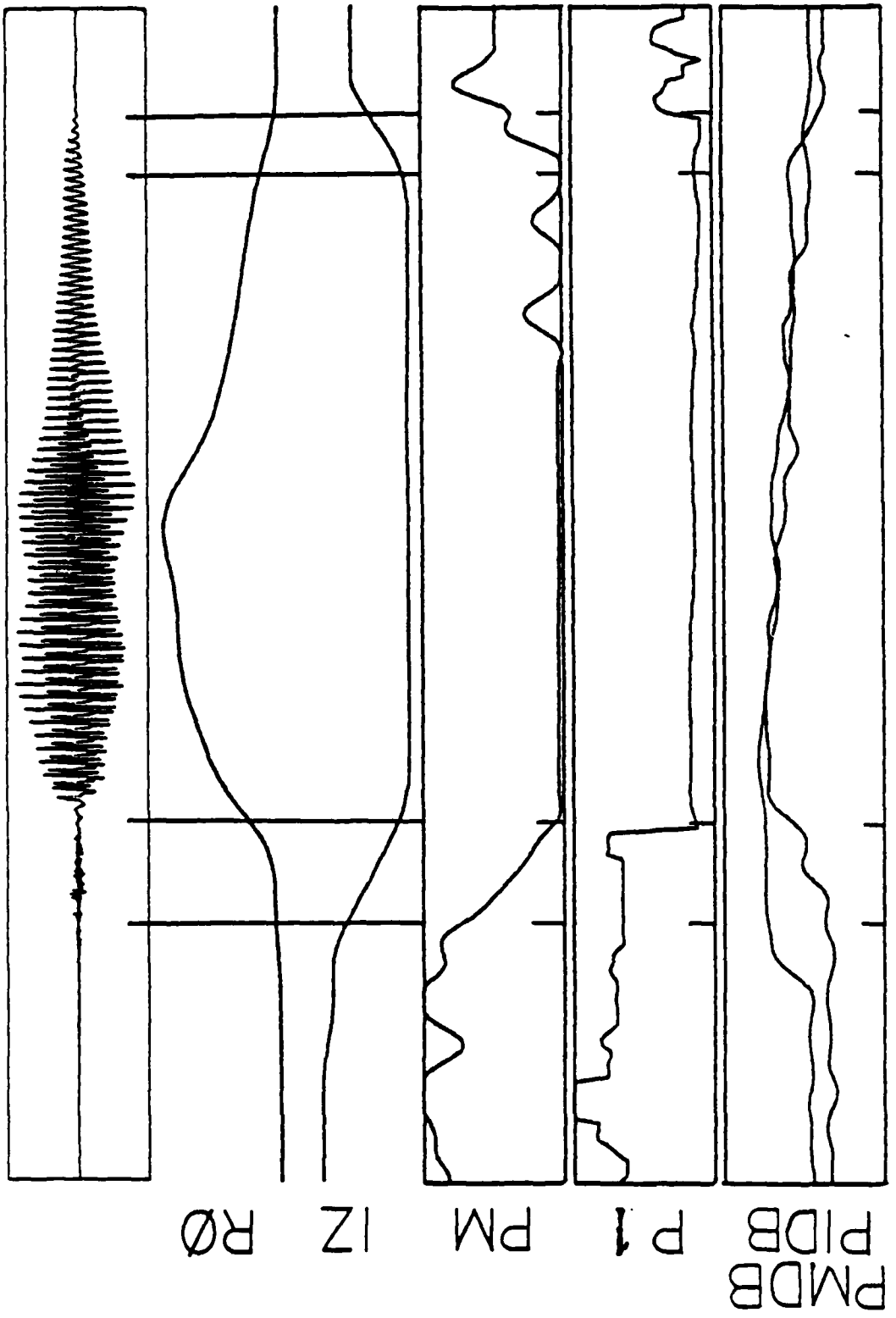


ONE (female: LT, wd7023.)

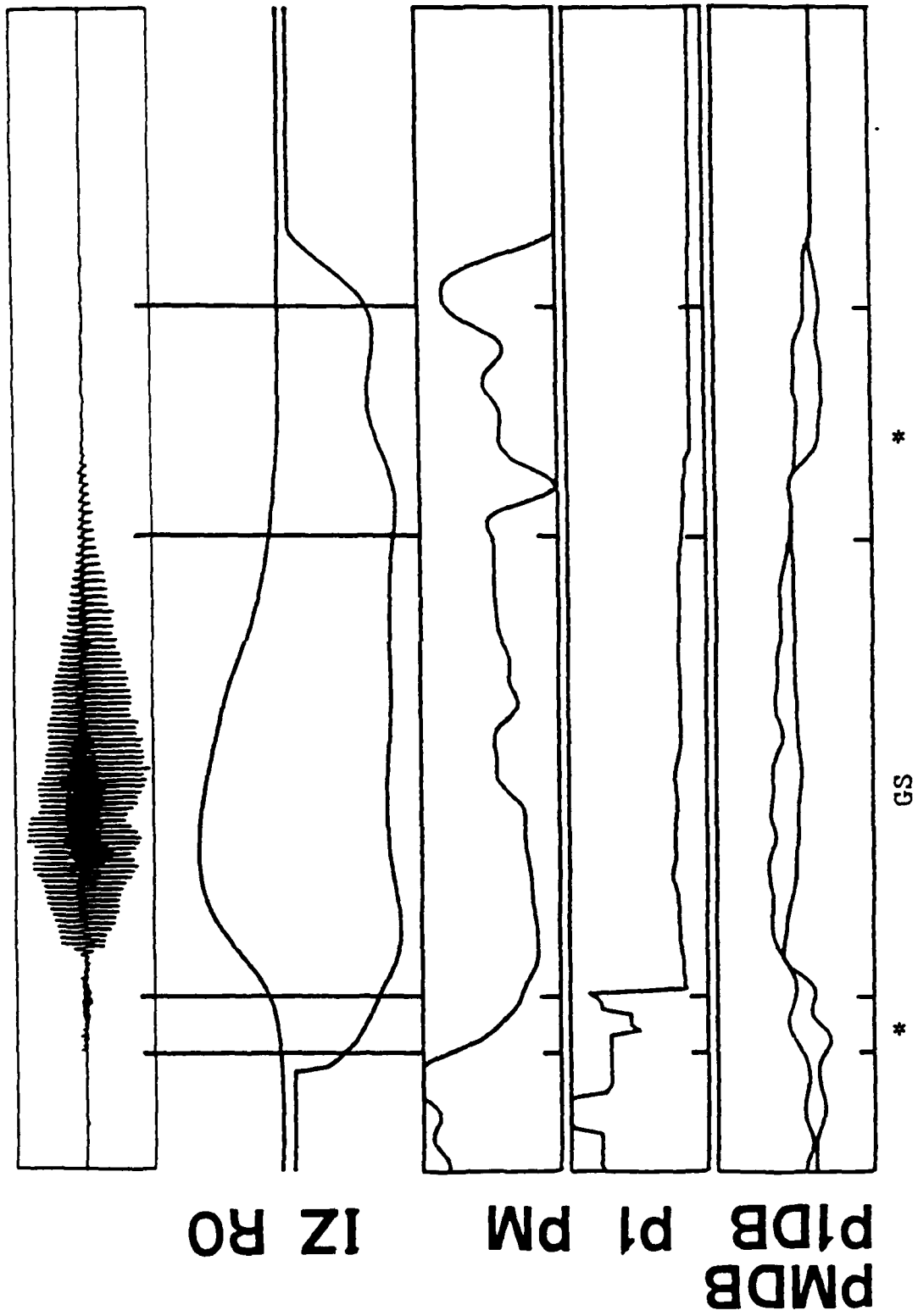

PMDB
PlDB
Pl
PM
lZ
RØ
GS
TWO
(male: MO, wd7002.)



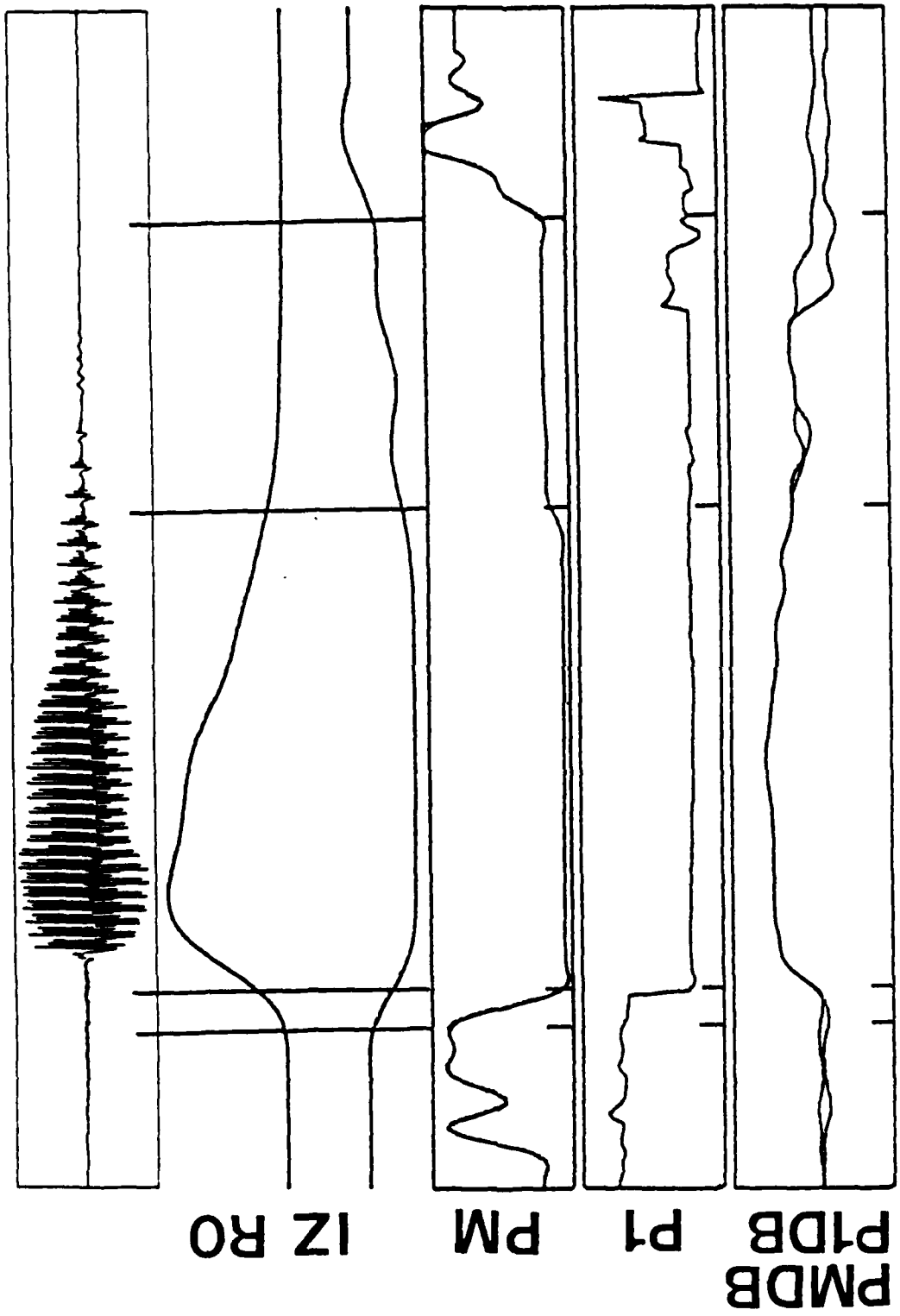
TWO (female: LT, wd7024.)



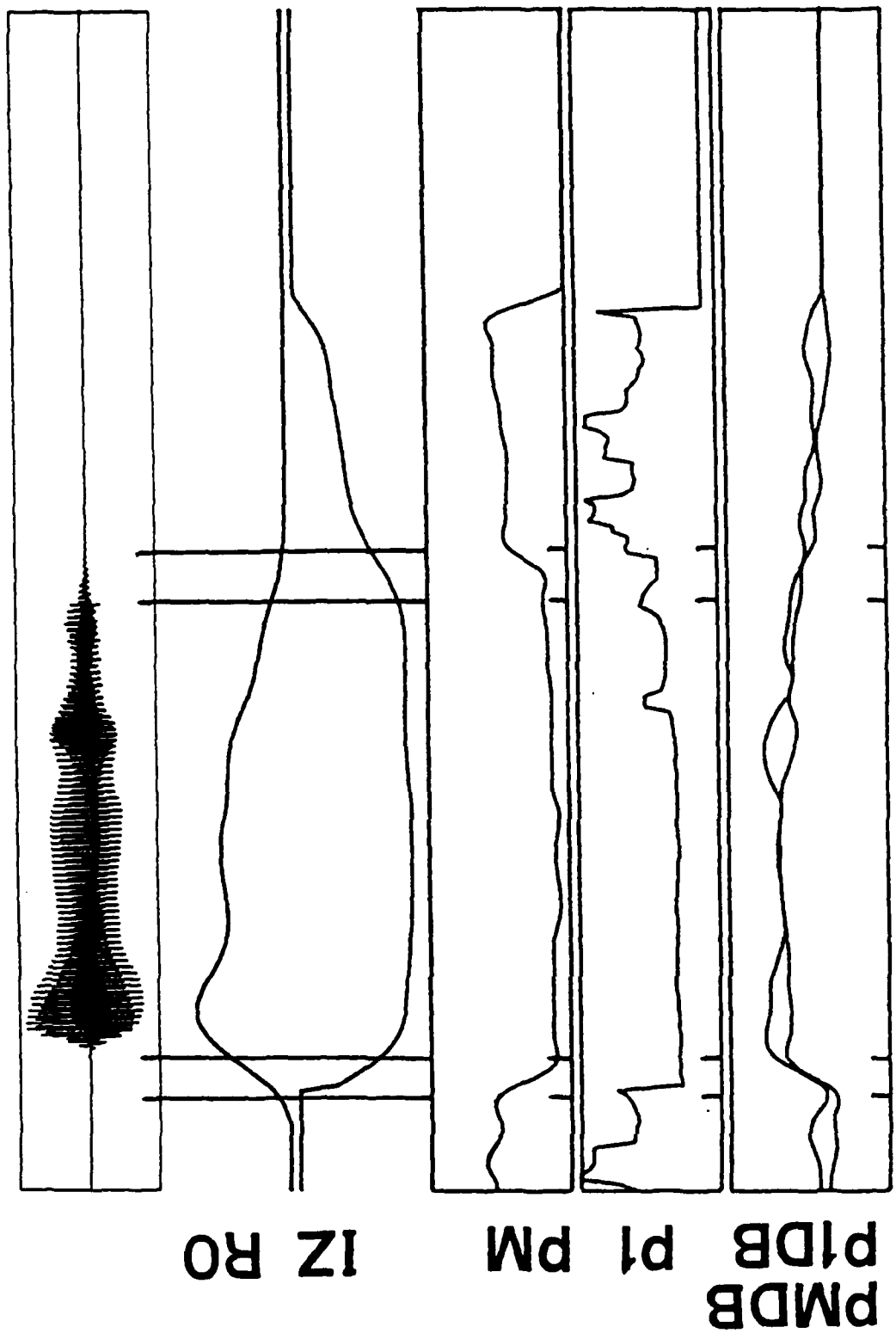
THREE (male: MO, wd7003.)



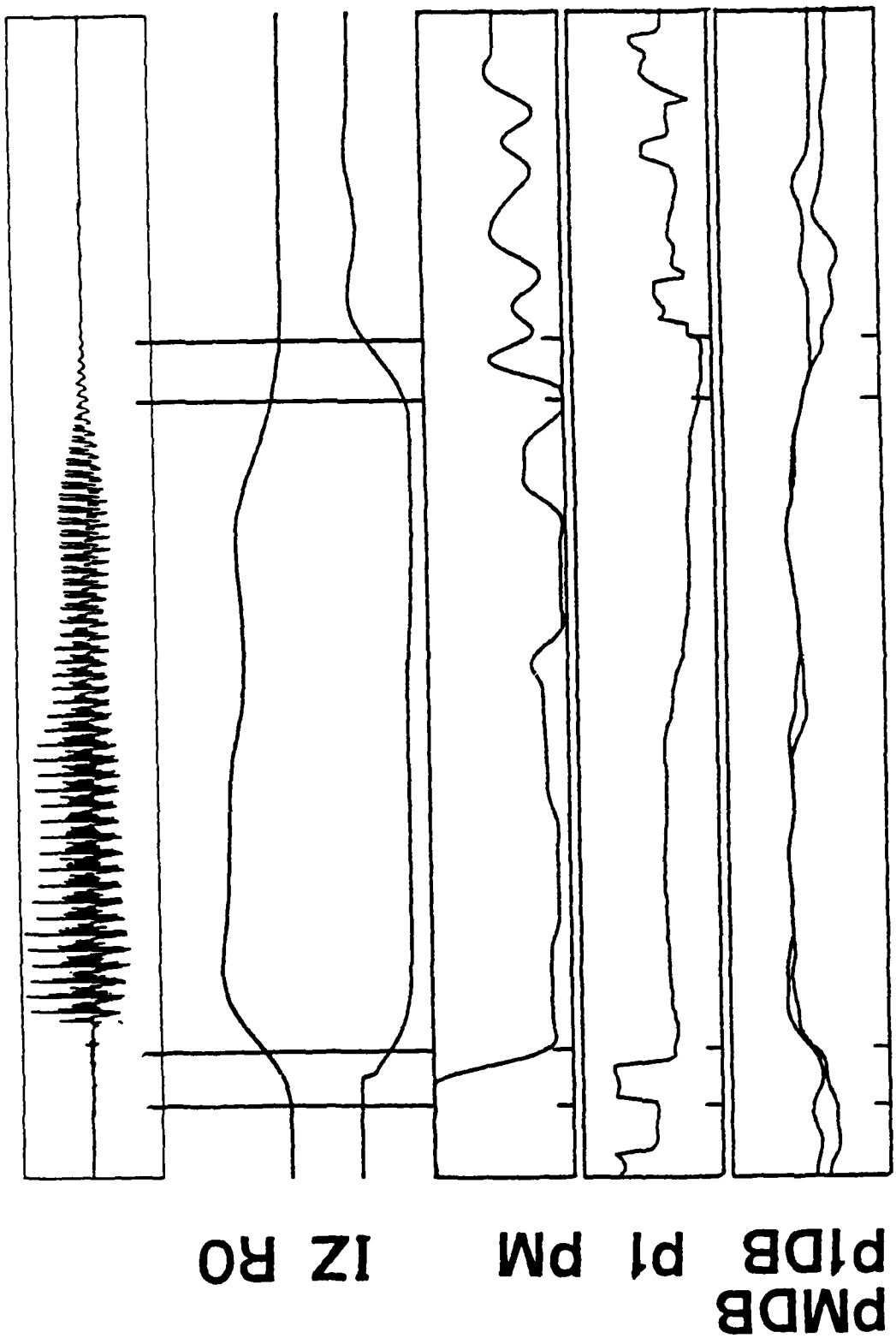
THREE (female: LT, wa7025.)



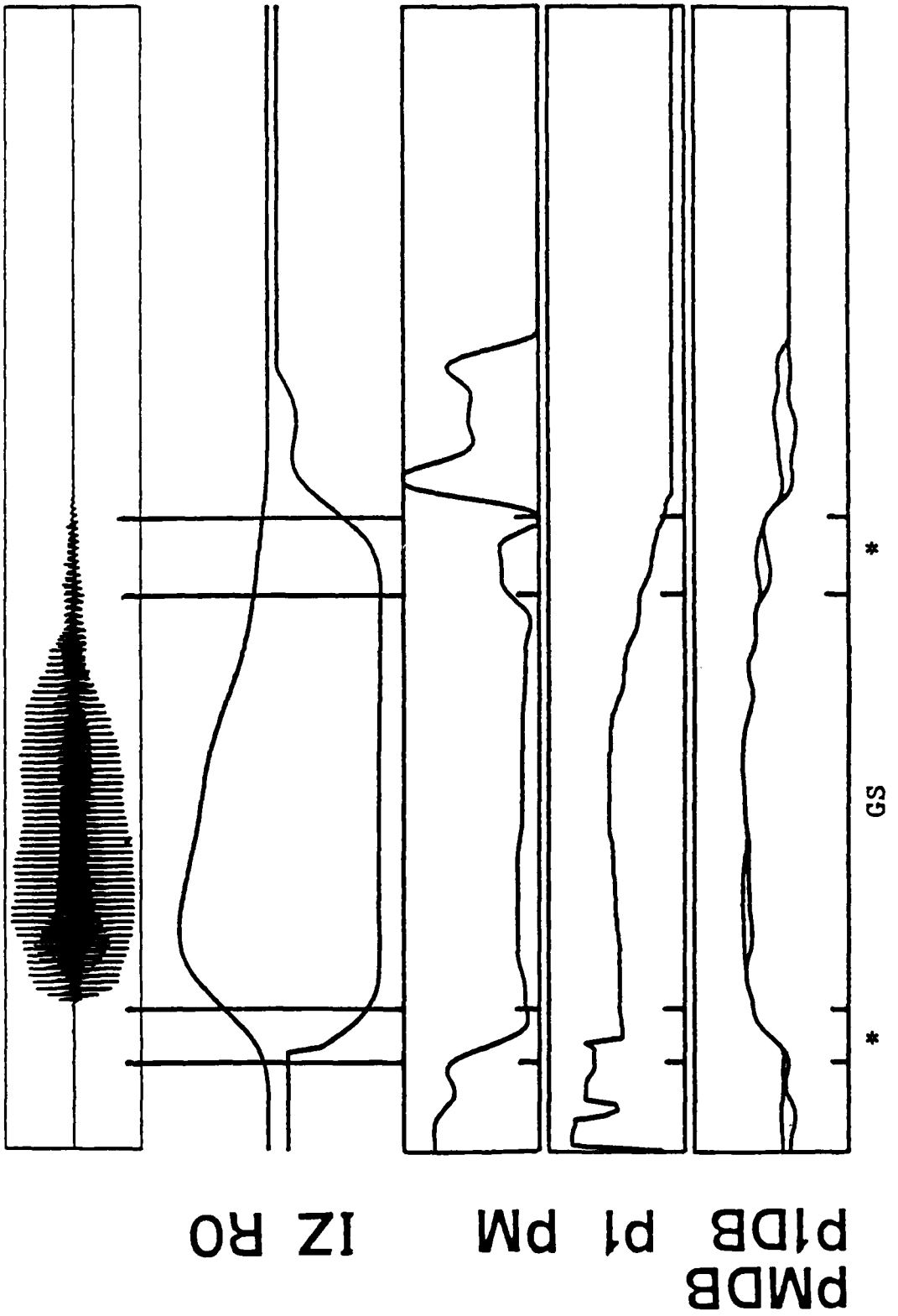
FOUR (male: MO, wd7004.)



FOUR (female: LT, wd7026.)

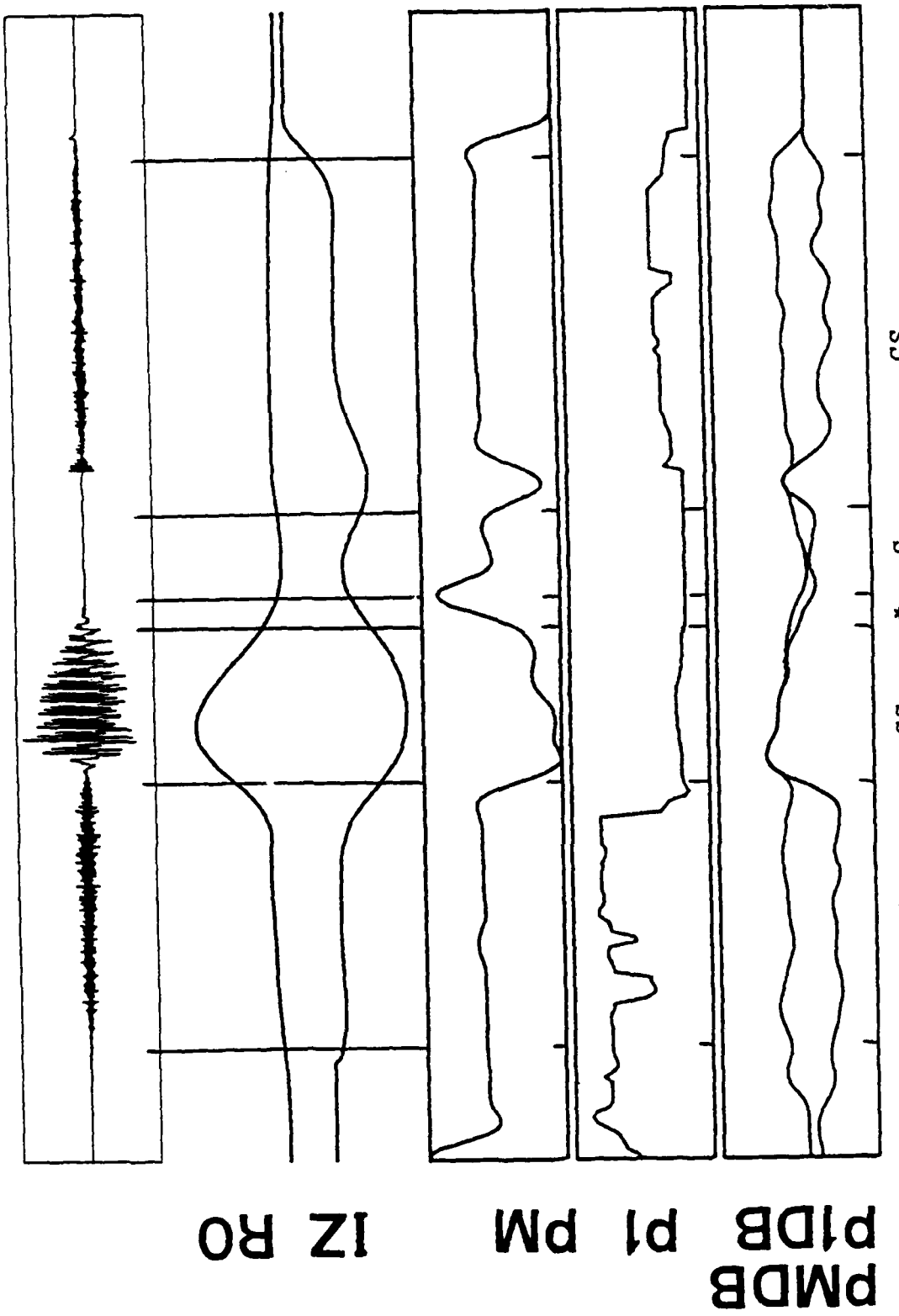


FIVE (male: MO, wd7005.)

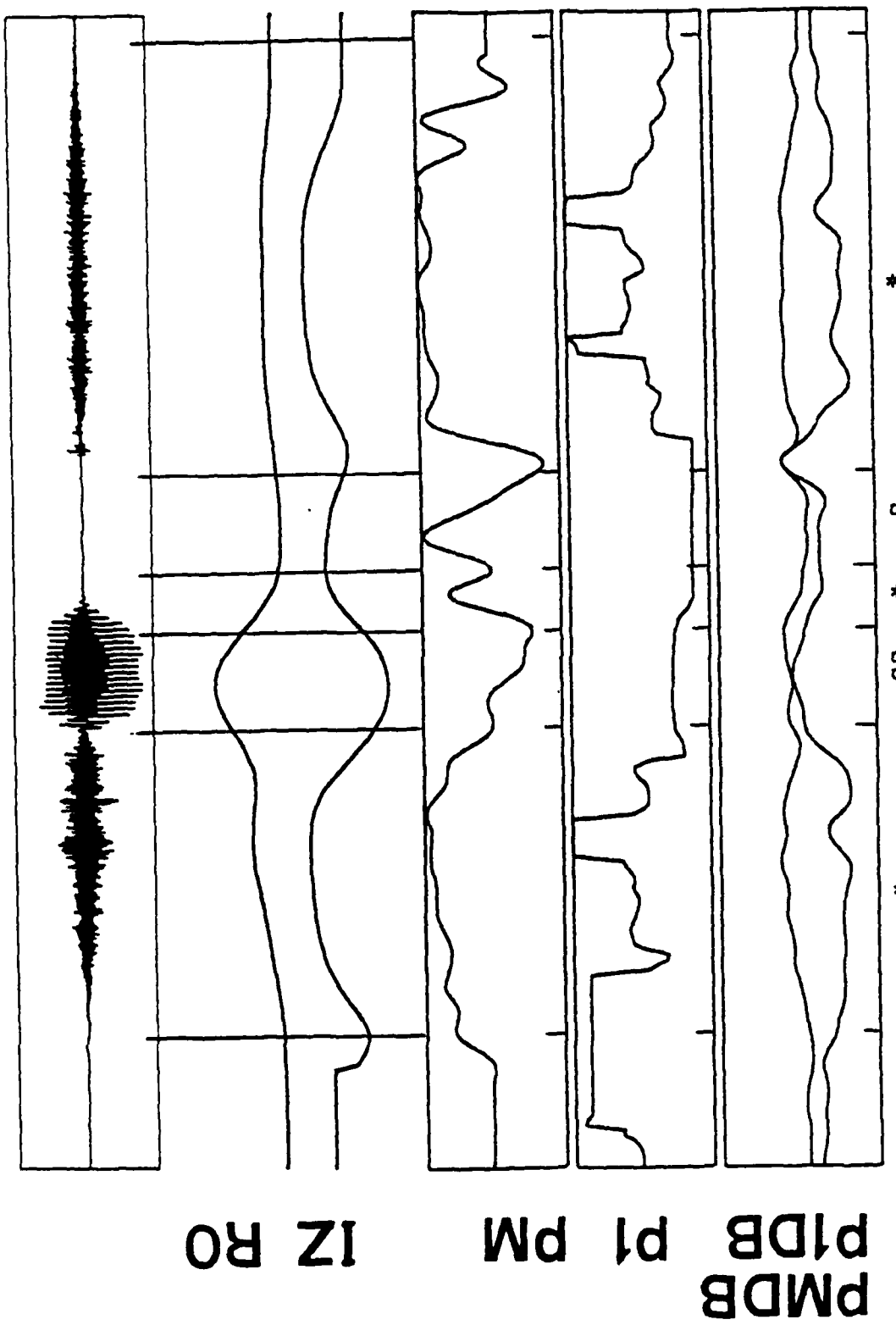


FIVE

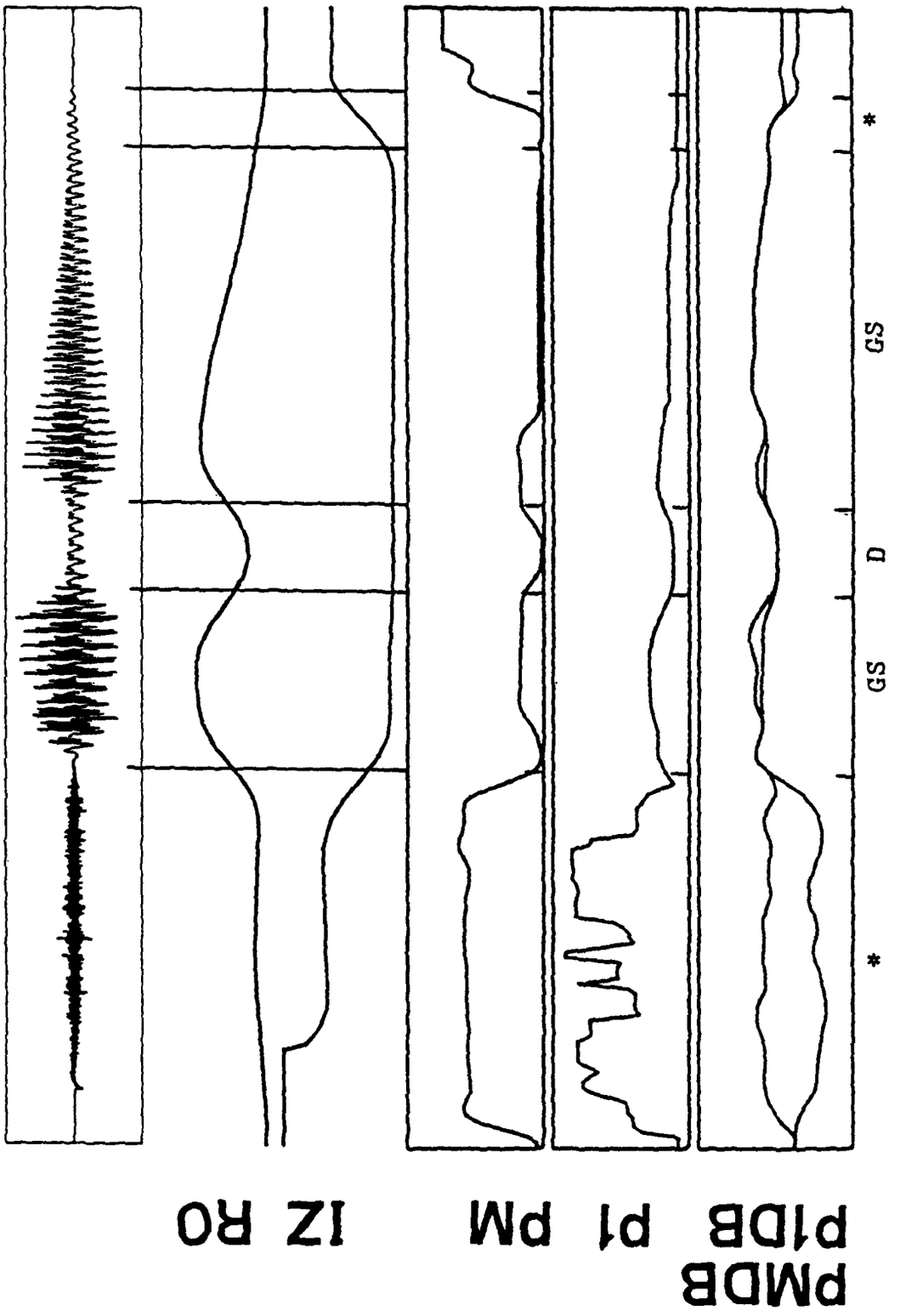
(female: LT, wd7027.)



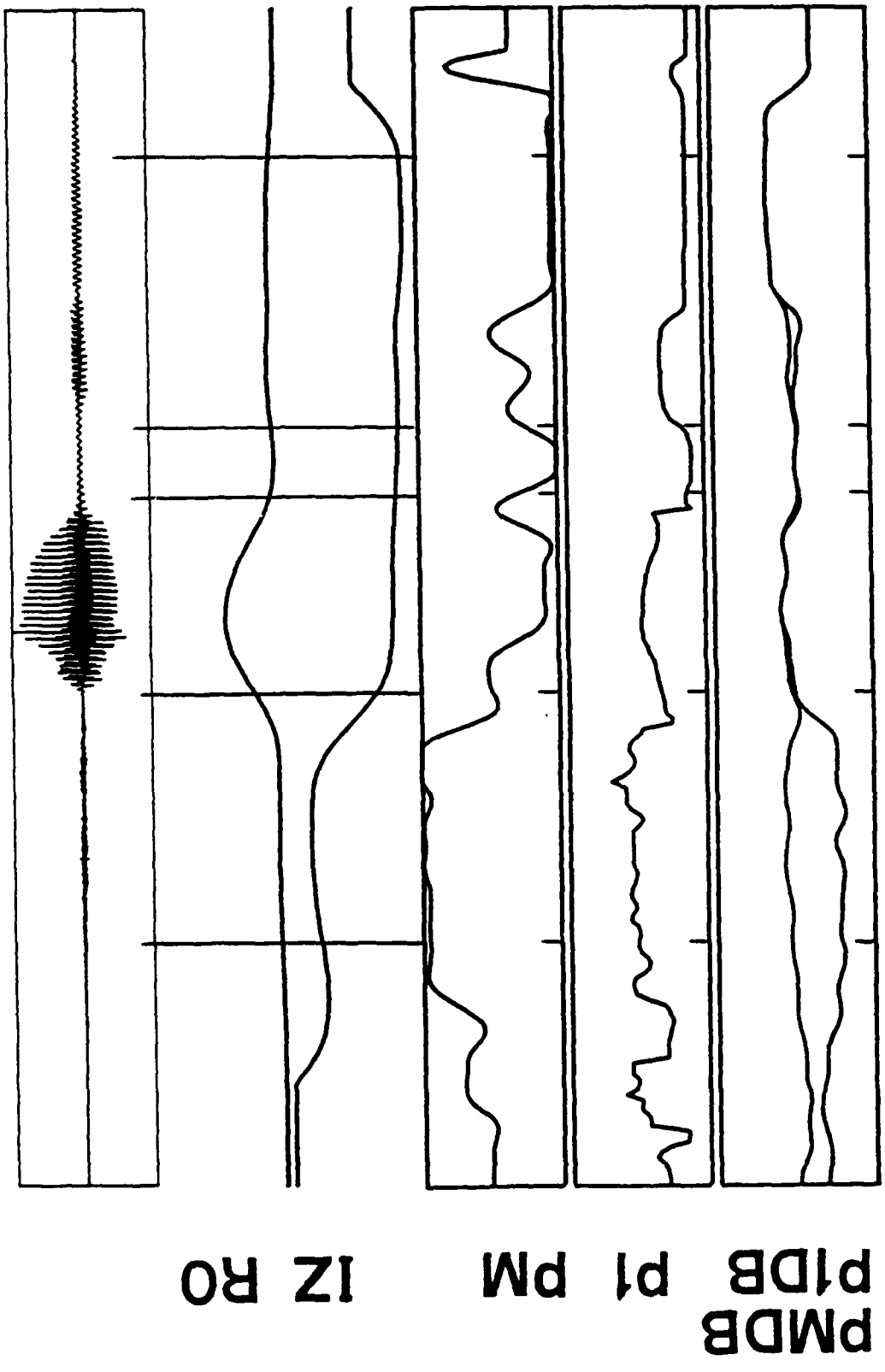
SIX (male: MO, wd7006.)



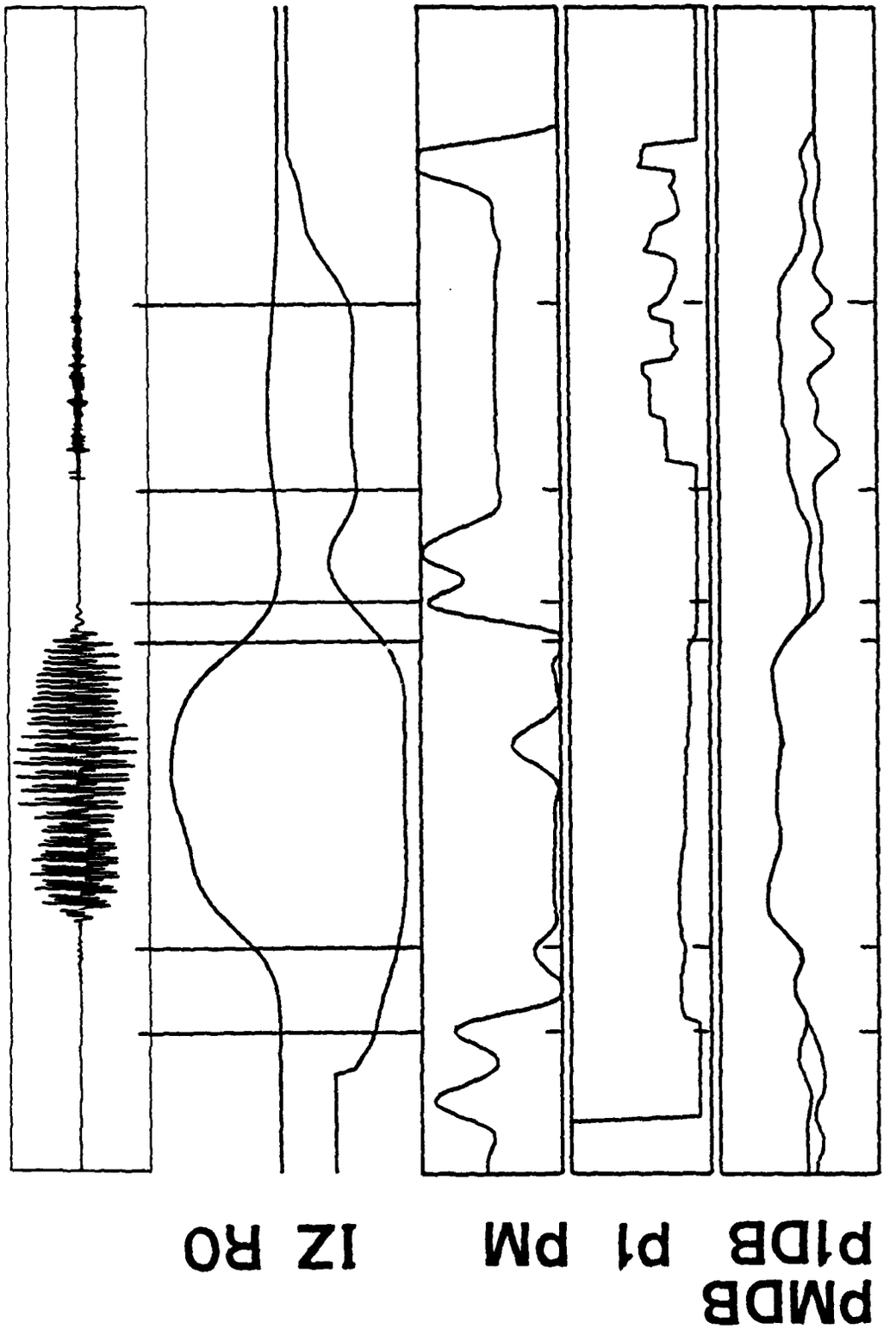
SIX (female: LT, wd7028.)



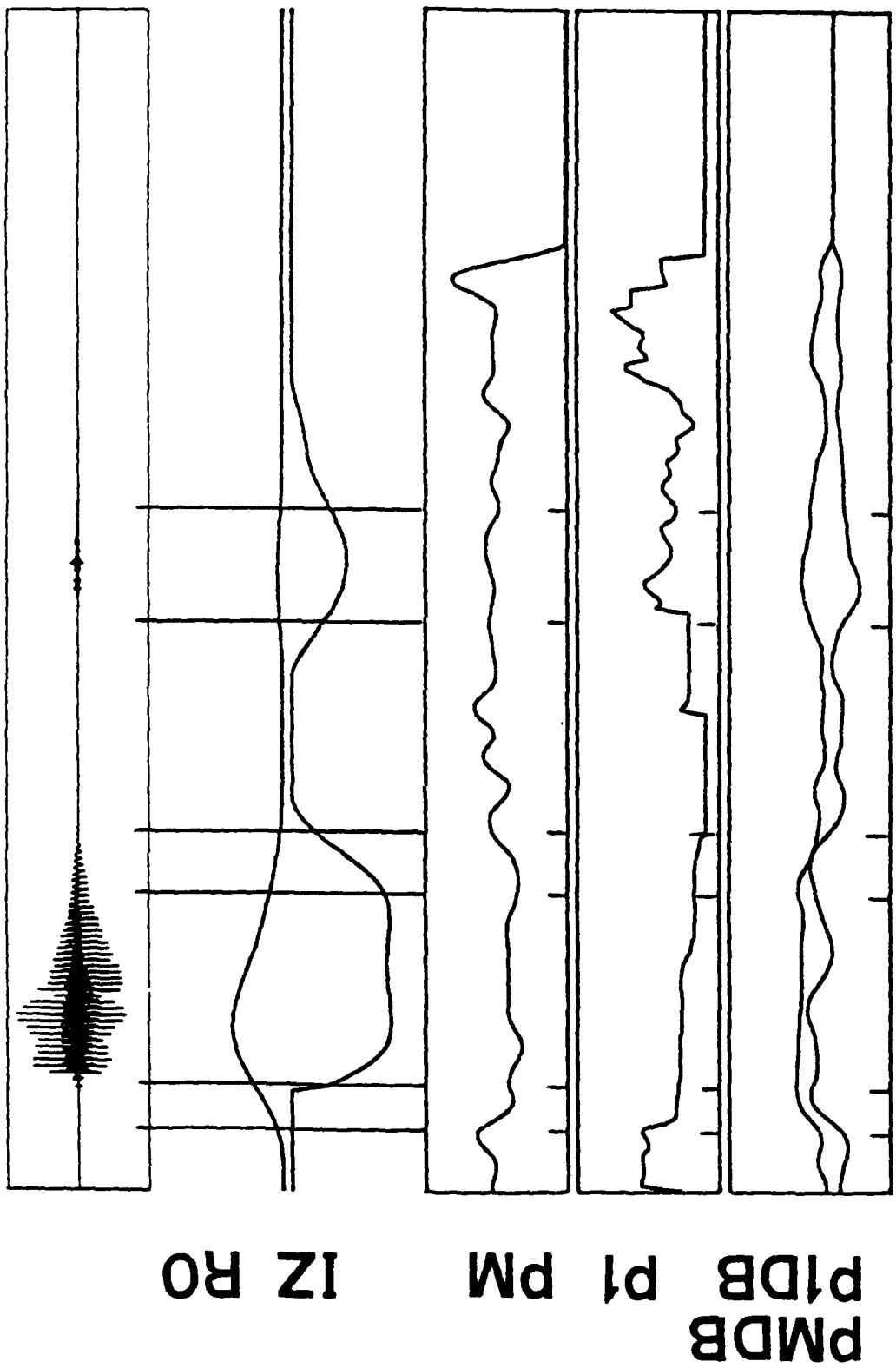
SEVEN (male: MO, wd7007.)



SEVEN (female: LT, wd7029.)

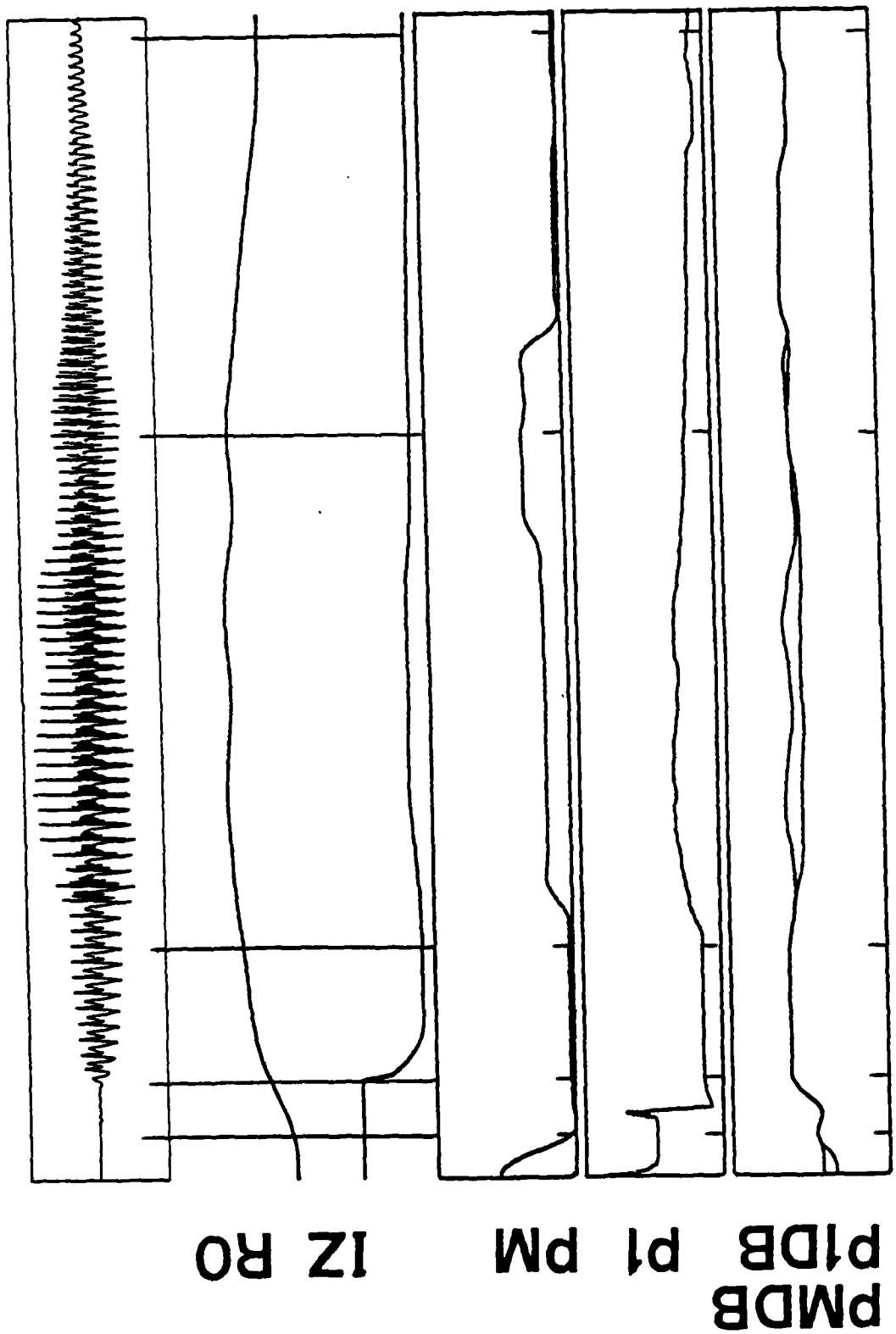


EIGHT (male: MO, wd7008.)

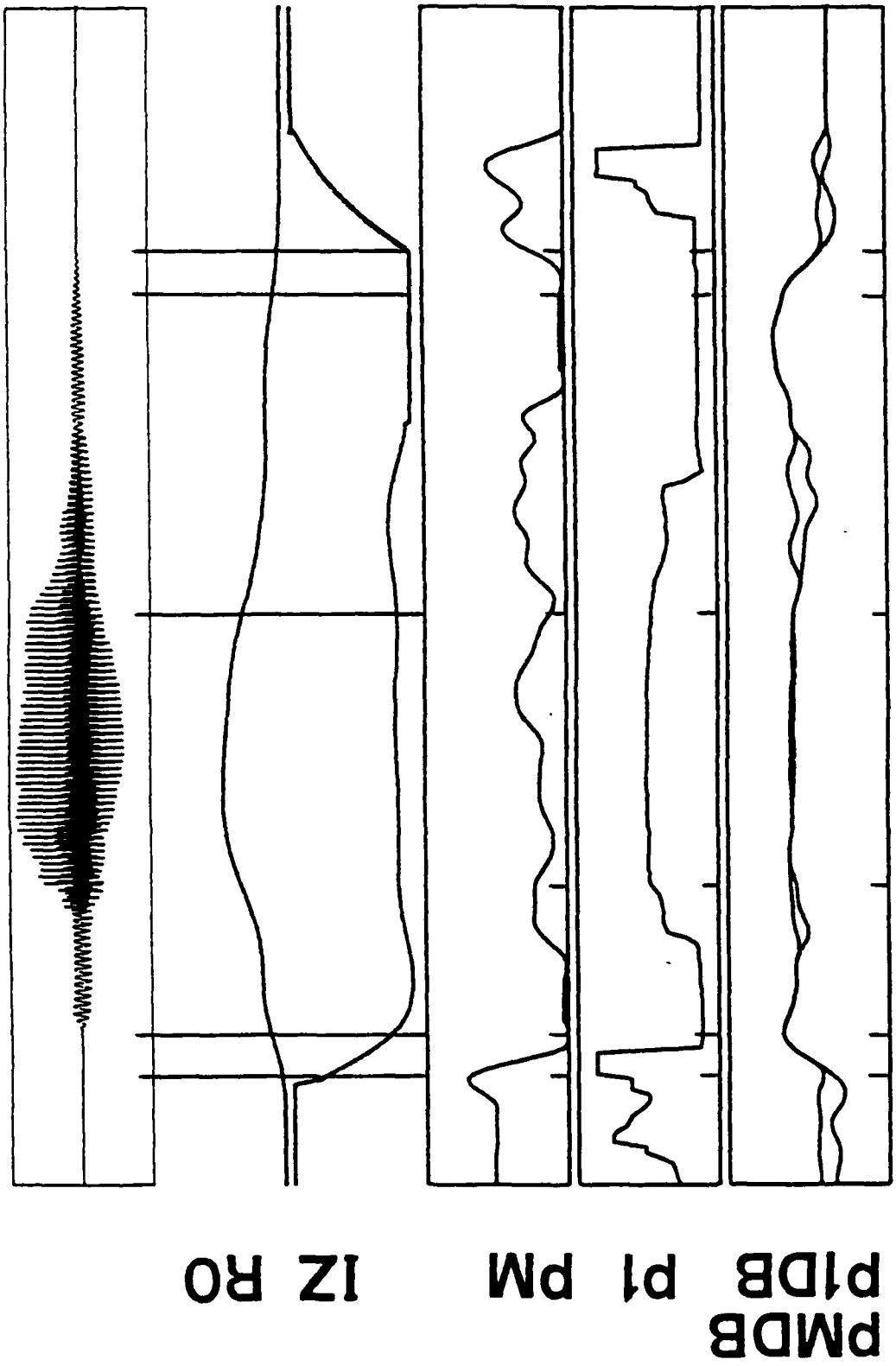


EIGHT

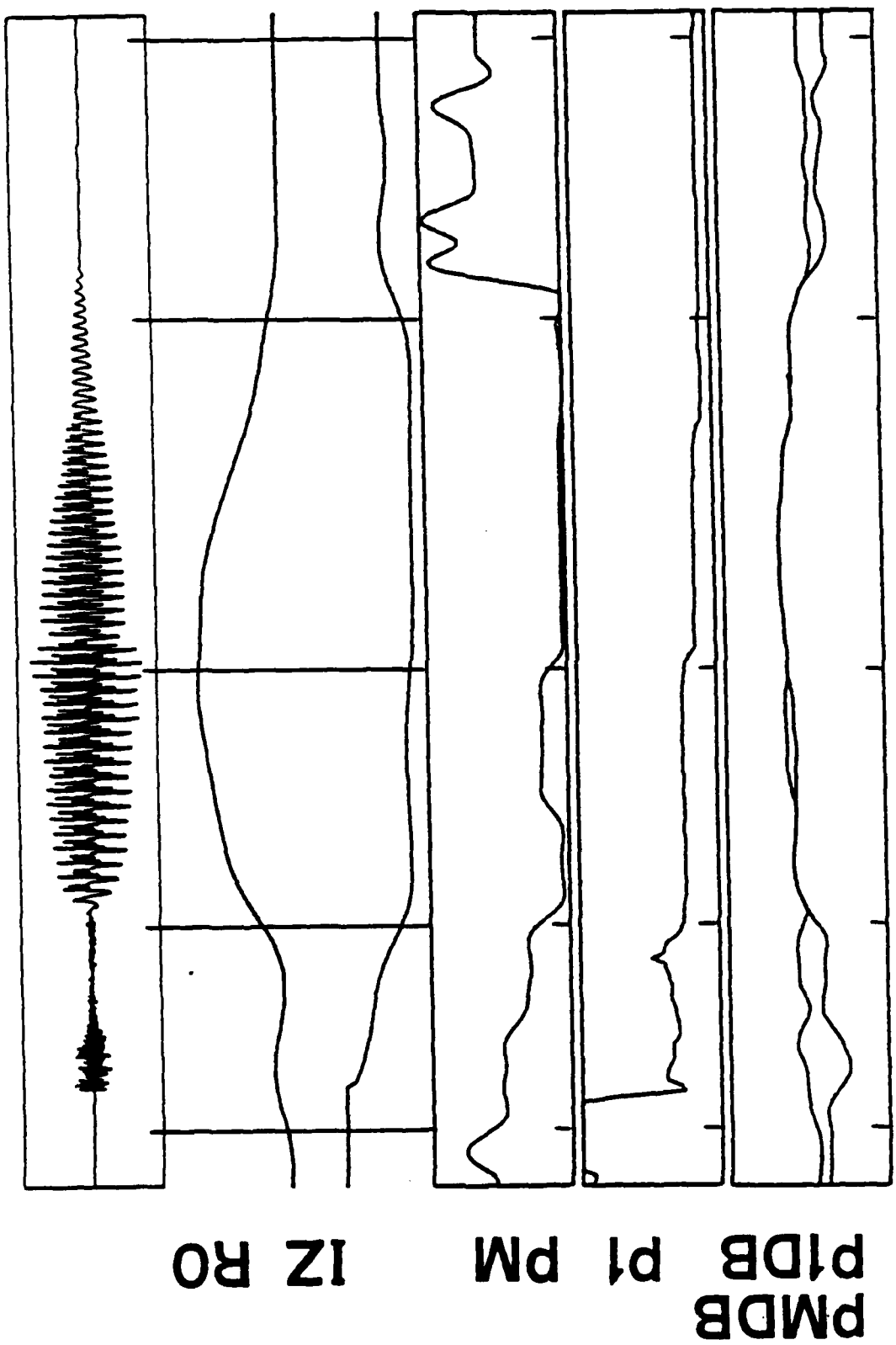
(female: LT, wd7030.)

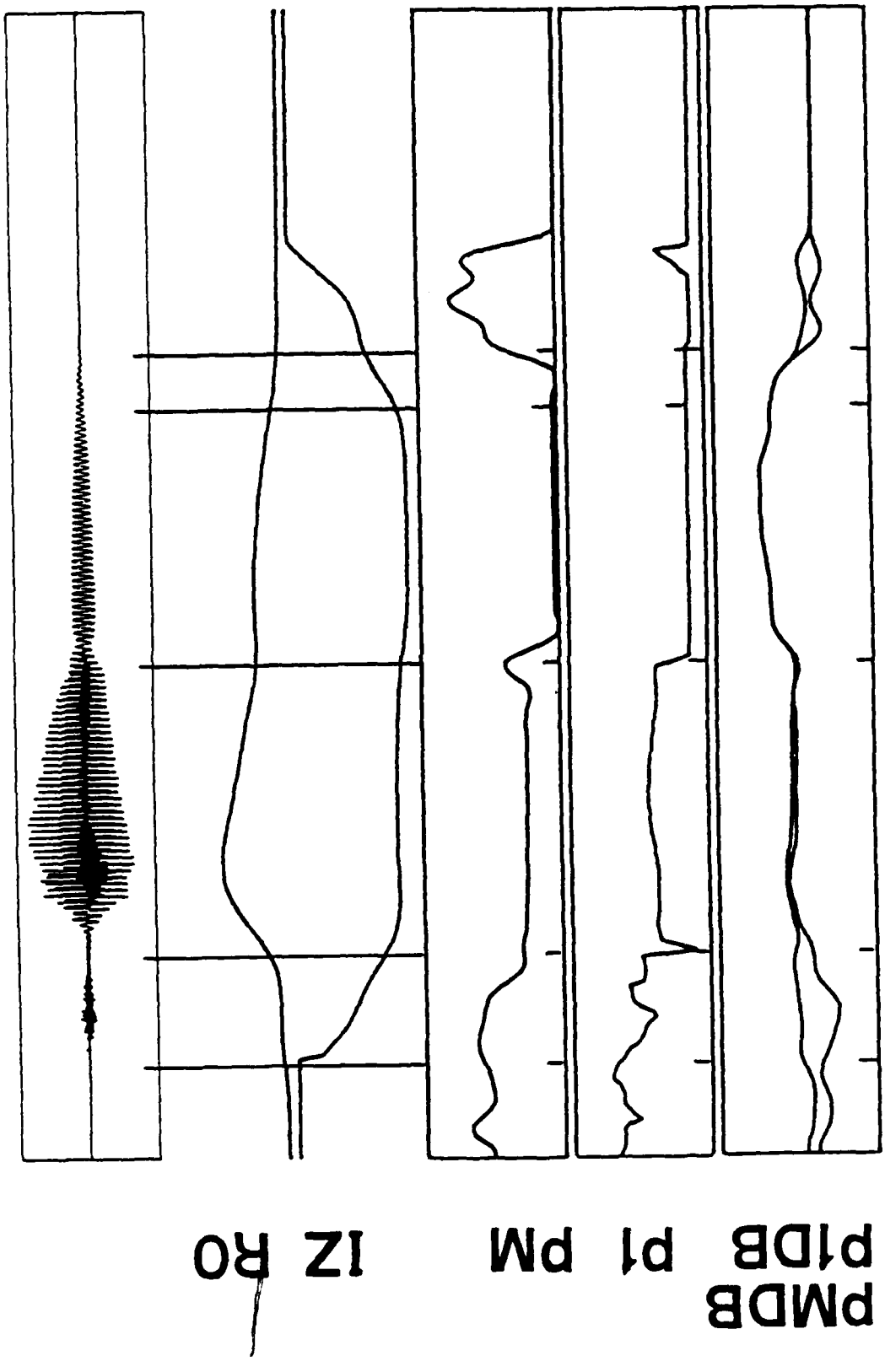


NINE (male: MO, wd7009.)



NINE (female: LT, wd7031.)


pMDB
p1DB
p1
pM
IZ
RO
TEN
(male: MO, wd7010.)
GS
N
*
*



TEN (female: LT, wd?032.)

APPENDIX 10.3

Listing of Sensory Formant Generation Errors
Output from an Error Analysis Program

The following listings are generated from an error analysis program called FMT_ERROR which reads a SWIS generated sensory formant data file and an expert-created sensory formant data file, where both files associated with the same utterance are compared frame by frame by FMT_ERROR program. The errors are output in the order of formants, i.e., the first output line corresponds to the first-sensory-formant-low (SF1L), the second to the first-sensory-formant-high (SF1H), the third to the second sensory formant, and the fourth to the third sensory formant.

The notations used in the error listings are explained below:

MIN:
the minimum difference in Hz between the two data files for a sensory formant;

MAX:
the maximum difference in Hz between the two data files for a sensory formant;

MEAN:
the mean of the absolute difference in Hz between the two data files for a sensory formant;

S.D.:
the standard derivation of the absolute difference in Hz between the two data files for a sensory formant;

MEAN (%):
the mean of the relative errors between the two data files for a sensory formant.

```
$DIG
$RUN HBOX:FMT_ERROR

                                "ZERO"  :  male utterance

 Enter first formant file name:
FMT9000.
 Enter second formant file name:
WD9000.FMT
     File Name (expert):FMT9000.          Frame range(first,last):   62  206
     File Name (SWIS)  :WD9000.FMT        Frame range(first,last):   58  193
     Frames compared :   133
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      0.0   241.0     24.917     23.689      0.078
      0.0   241.0     24.917     23.689      0.078
      1.0   878.0     61.835     84.157      0.064
      1.0  2334.0     63.872    209.626      0.040
$RUN HBOX:FMT_ERROR

                                "ONE"

 Enter first formant file name:
FMT9001.
 Enter second formant file name:
WD9001.FMT
     File Name (expert):FMT9001.          Frame range(first,last):   37  120
     File Name (SWIS)  :WD9001.FMT        Frame range(first,last):   36  203
     Frames compared :    84
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      0.0    58.0      22.798     15.364      0.060
      0.0    58.0      22.798     15.364      0.060
     14.0   553.0     161.643    159.646      0.264
      2.0   159.0      35.786     34.926      0.014
$RUN HBOX:FMT_ERROR
                                "TWO"

 Enter first formant file name:
FMT9002.
 Enter second formant file name:
WD9002.FMT
     File Name (expert):FMT9002.          Frame range(first,last):   70  186
     File Name (SWIS)  :WD9002.FMT        Frame range(first,last):   64  181
     Frames compared :   112
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      1.0   125.0     29.116     23.156      0.102
      1.0   125.0     29.116     23.156      0.102
      0.0   186.0     46.982     36.973      0.042
      1.0   359.0     40.179     65.820      0.018
```

```
$RUN HBOX:FMT_ERROR

                              "THREE"

 Enter first formant file name:
FMT9003.
 Enter second formant file name:
WD9003.FMT
     File Name (expert):FMT9003.        Frame range(first,last):   73  205
     File Name (SWIS)  :WD9003.FMT      Frame range(first,last):   69  199
     Frames compared :   127

     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      1.0     67.0     24.142     14.626      0.090
      1.0     67.0     24.142     14.626      0.090
      2.0    217.0     54.268     43.790      0.029
      1.0   1369.0    169.039    374.620      0.083
$RUN HBOX:FMT_ERROR

                              "FOUR"

 Enter first formant file name:
FMT9004.
 Enter second formant file name:
WD9004.FMT
     File Name (expert):FMT9004.        Frame range(first,last):   46  146
     File Name (SWIS)  :WD9004.FMT      Frame range(first,last):   39  134
     Frames compared :    89
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
     21.0     66.0     41.079     14.429      0.093
     21.0     66.0     41.079     14.429      0.093
      0.0    201.0     50.079     46.010      0.055
      1.0    641.0     99.685    142.284      0.052
$RUN HBOX:FMT_ERROR

                              "FIVE"

 Enter first formant file name:
FMT9005.
 Enter second formant file name:
WD9005.FMT
     File Name (expert):FMT9005.        Frame range(first,last):   32  144
     File Name (SWIS)  :WD9005.FMT      Frame range(first,last):   28  153
     Frames compared :   113
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      2.0     59.0     27.920     14.264      0.056
      2.0     59.0     27.920     14.264      0.056
      1.0    163.0     50.398     40.544      0.032
      0.0     61.0     19.044     12.713      0.008
```

```
$RUN HBOX:FMT_ERROR

                                "SIX"

 Enter first formant file name:
FMT9006.
 Enter second formant file name:
WD9006.FMT
     File Name (expert):FMT9006.        Frame range(first,last):   80  110
     File Name (SWIS)  :WD9006.FMT      Frame range(first,last):   77  104
     Frames compared :    25
     MIN      MAX       MEAN       S.D      MEAN(%)
     ------------------------------------------------
      4.0     50.0     27.160     15.445      0.070
      4.0     50.0     27.160     15.445      0.070
      4.0     54.0     35.560     13.599      0.021
      1.0   1131.0    100.640    244.341      0.039
$RUN HBOX:FMT_ERROR

                               "SEVEN"

 Enter first formant file name:
FMT9007.

 Enter second formant file name:
WD9007.FMT
     File Name (expert):FMT9007.        Frame range(first,last):   79  113
     File Name (SWIS)  :WD9007.FMT      Frame range(first,last):   76  114
     Frames compared :    35
     MIN      MAX       MEAN       S.D      MEAN(%)
     ------------------------------------------------
      2.0     40.0     22.543     12.908      0.045
      2.0     40.0     22.543     12.908      0.045
      2.0     83.0     39.457     26.559      0.027
      2.0     35.0     13.286      7.733      0.005
$RUN HBOX:FMT_ERROR

                               "EIGHT"

 Enter first formant file name:
FMT9008.
 Enter second formant file name:
WD9008.FMT
     File Name (expert):FMT9008.        Frame range(first,last):   50  109
     File Name (SWIS)  :WD9008.FMT      Frame range(first,last):   45  105
     Frames compared :    56
     MIN      MAX       MEAN       S.D      MEAN(%)
     ------------------------------------------------
      0.0     73.0     26.893     13.024      0.079
      0.0     73.0     26.893     13.024      0.079
      0.0    425.0     72.875     85.486      0.032
      1.0   1450.0    137.821    301.381      0.047
```

```
$RUN HBOX:FMT_ERROR

                        "NINE"

 Enter first formant file name:
FMT9009.
 Enter second formant file name:
WD9009.FMT
     File Name (expert):FMT9009.        Frame range(first,last):   57  164
     File Name (SWIS)  :WD9009.FMT      Frame range(first,last):   24  225
     Frames compared :   108
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      0.0     43.0      18.019     10.646      0.039
      0.0     43.0      18.019     10.646      0.039
      0.0    204.0      43.843     42.160      0.025
      0.0    127.0      30.102     24.085      0.012
$RUN HBOX:FMT_ERROR
                        "ZERO"  :  female utterance


 Enter first formant file name:
FMT9022.
 Enter second formant file name:
WD9022.FMT
     File Name (expert):FMT9022.        Frame range(first,last):   67  169
     File Name (SWIS)  :WD9022.FMT      Frame range(first,last):   61  171
     Frames compared :   103
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      1.0    122.0      34.320     23.687      0.054
      1.0    122.0      34.320     23.687      0.054
      4.0    178.0      62.903     45.334      0.044
      0.0    322.0      57.427     68.898      0.029
$RUN HBOX:FMT_ERROR
                        "ONE"

 Enter first formant file name:
FMT9023.
 Enter second formant file name:
WD9023.FMT
     File Name (expert):FMT9023.        Frame range(first,last):   29   86
     File Name (SWIS)  :WD9023.FMT      Frame range(first,last):   29  132
     Frames compared :    58
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      0.0    151.0      23.310     25.839      0.034
      0.0    142.0      23.155     25.074      0.032
      1.0    230.0      93.914     68.818      0.069
      0.0    102.0      31.724     24.500      0.010
```

```
$RUN HBOX:FMT_ERROR
                              "TWO"


 Enter first formant file name:
FMT9024.
 Enter second formant file name:
WD9024.FMT
     File Name (expert):FMT9024.        Frame range(first,last):   57  151
     File Name (SWIS)  :WD9024.FMT      Frame range(first,last):   53  143
     Frames compared :    87
     MIN      MAX       MEAN       S.D      MEAN(%)
     ------------------------------------------------
      1.0     32.0      14.529      5.949      0.046
      1.0     32.0      14.529      5.949      0.046
      4.0    467.0     115.333    121.879      0.090
      0.0     88.0      24.471     21.481      0.009
$RUN HBOX:FMT_ERROR
                             "THREE"


 Enter first formant file name:
FMT9025.
 Enter second formant file name:
WD9025.FMT
     File Name (expert):FMT9025.        Frame range(first,last):   45  142
     File Name (SWIS)  :WD9025.FMT      Frame range(first,last):   38  134
     Frames compared :    90
     MIN      MAX       MEAN       S.D      MEAN(%)
     ------------------------------------------------
      0.0     47.0      18.933     13.790      0.051
      0.0     47.0      18.933     13.790      0.051
      1.0    221.0      54.122     50.322      0.025
      2.0    459.0     162.078    123.908      0.058
$RUN HBOX:FMT_ERROR
                              "FOUR"


 Enter first formant file name:
FMT9026.
 Enter second formant file name:
WD9026.FMT
     File Name (expert):FMT9026.        Frame range(first,last):   30  124
     File Name (SWIS)  :WD9026.FMT      Frame range(first,last):   28  114
     Frames compared :    85
     MIN      MAX       MEAN       S.D      MEAN(%)
     ------------------------------------------------
     5.0     134.0      41.071     30.750      0.062
     5.0     134.0      41.071     30.750      0.062
     0.0     185.0      50.765     36.547      0.048
     2.0     273.0      49.976     58.360      0.024
```

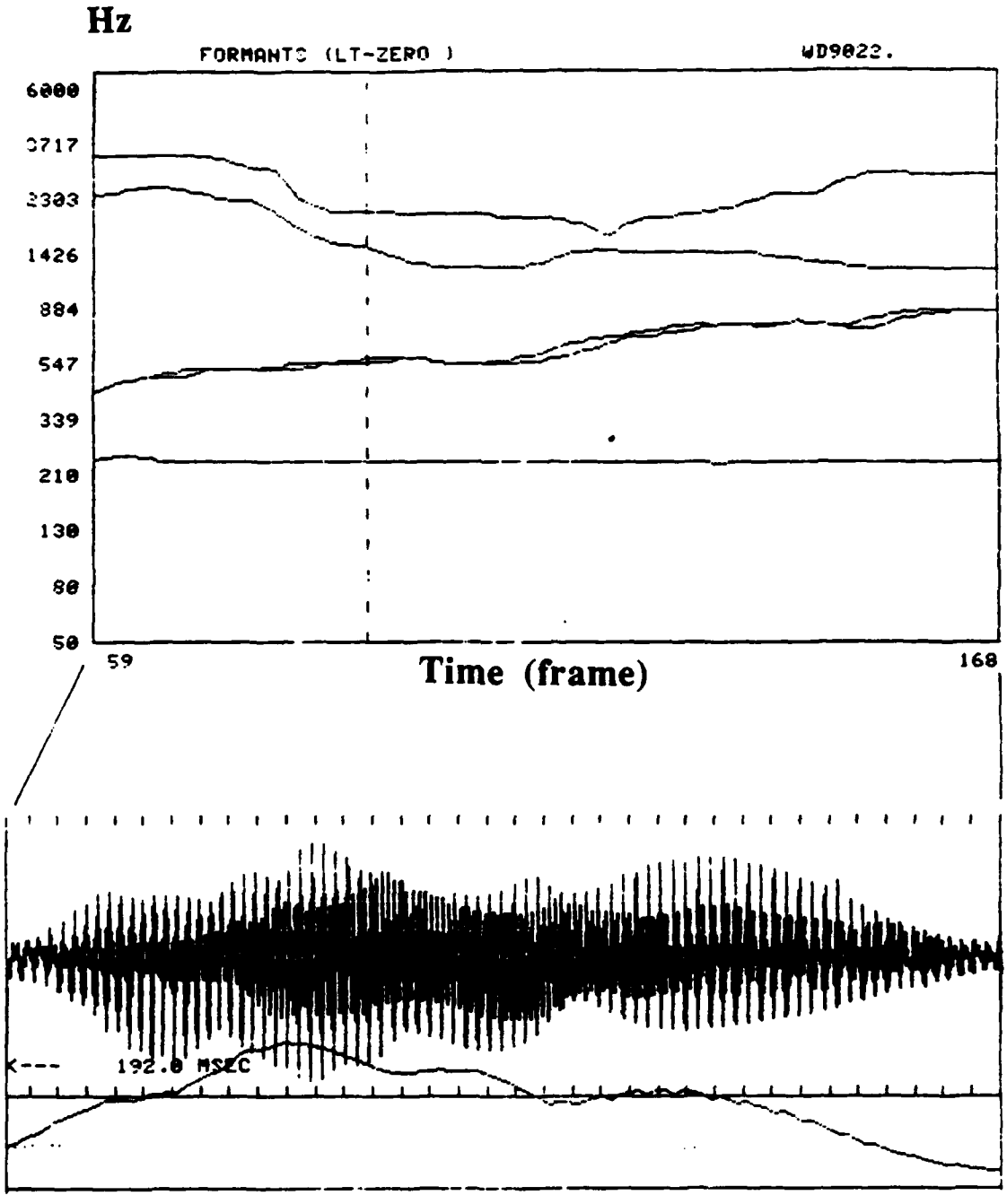
```
$RUN HBOX:FMT_ERROR

                              "FIVE"


 Enter first formant file name:
FMT9027.
 Enter second formant file name:
WD9027.FMT
     File Name (expert):FMT9027.        Frame range(first,last):   31  104
     File Name (SWIS)  :WD9027.FMT      Frame range(first,last):   30  121
     Frames compared :    74
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      1.0    147.0     67.797     52.665      0.067
      1.0    147.0     67.797     52.665      0.067
      2.0    241.0     63.216     60.209      0.036
      0.0     55.0     14.595     13.190      0.005
$RUN HBOX:FMT_ERROR

                              "SIX"


 Enter first formant file name:
FMT9028.
 Enter second formant file name:
WD9028.FMT
     File Name (expert):FMT9028.        Frame range(first,last):   86  113
     File Name (SWIS)  :WD9028.FMT      Frame range(first,last):   90  109
     Frames compared :    20
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      5.0     42.0     26.000     10.402      0.052
      5.0     42.0     26.000     10.402      0.052
      4.0     84.0     30.150     27.103      0.013
      4.0     82.0     37.850     23.100      0.013
$RUN HBOX:FMT_ERROR

                             "SEVEN"


 Enter first formant file name:
FMT9029.
 Enter second formant file name:
WD9029.FMT
     File Name (expert):FMT9029.        Frame range(first,last):   96  131
     File Name (SWIS)  :WD9029.FMT      Frame range(first,last):   97  135
     Frames compared :    35
     MIN      MAX       MEAN       S.D       MEAN(%)
     ------------------------------------------------
      6.0    216.0     38.171     36.914      0.049
      6.0    216.0     38.171     36.914      0.049
      1.0    179.0     68.686     56.315      0.041
      1.0     78.0     22.743     17.525      0.008
```

```
$RUN HBOX:FMT_ERROR

                              "EIGHT"

 Enter first formant file name:
FMT9030.
 Enter second formant file name:
WD9030.FMT

     File Name (expert):FMT9030.        Frame range(first,last):   22   68
     File Name (SWIS)  :WD9030.FMT      Frame range(first,last):   23   60
     Frames compared :    38
     MIN      MAX       MEAN       S.D      MEAN(%)
     ------------------------------------------------
      2.0     55.0     25.500     13.912      0.058
      2.0     55.0     25.500     13.912      0.058
      4.0   1804.0    469.763    682.600      0.217
      7.0    626.0    334.263    201.342      0.101
$RUN HBOX:FMT_ERROR

                              "NINE"

 Enter first formant file name:
FMT9031.
 Enter second formant file name:
WD9031.FMT
     File Name (expert):FMT9031.        Frame range(first,last):   56   144
     File Name (SWIS)  :WD9031.FMT      Frame range(first,last):   31   176
     Frames compared :    89
     MIN      MAX       MEAN       S.D      MEAN(%)
     ------------------------------------------------
      1.0    447.0     42.562     79.673      0.063
      1.0    125.0     28.517     26.043      0.034
      3.0    670.0    111.798    125.292      0.054
      0.0    478.0     61.404    114.718      0.020
$EXIT

  HMC          job terminated at 28-AUG-1986 16:48:27.33

  Accounting information:
  Buffered I/O count:          600      Peak working set size:    984
  Direct I/O count:            959      Peak page file size:     1812
  Page faults:                6252      Mounted volumes:            0
  Charged CPU time:     0 00:01:04.33   Elapsed time:     0 00:02:36.52
```

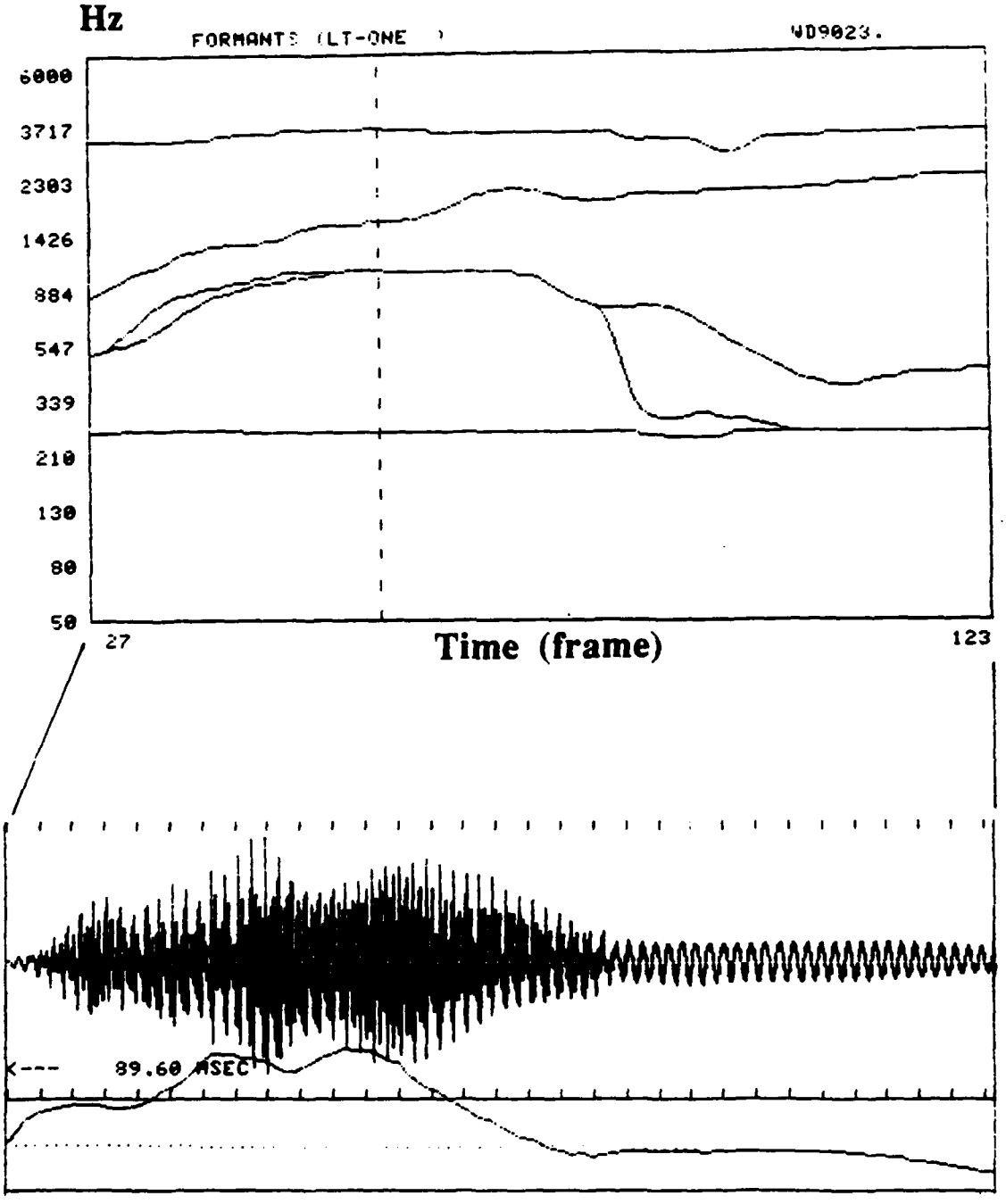
APPENDIX 10.4

Sample Plots of Smoothed Sensory Formants

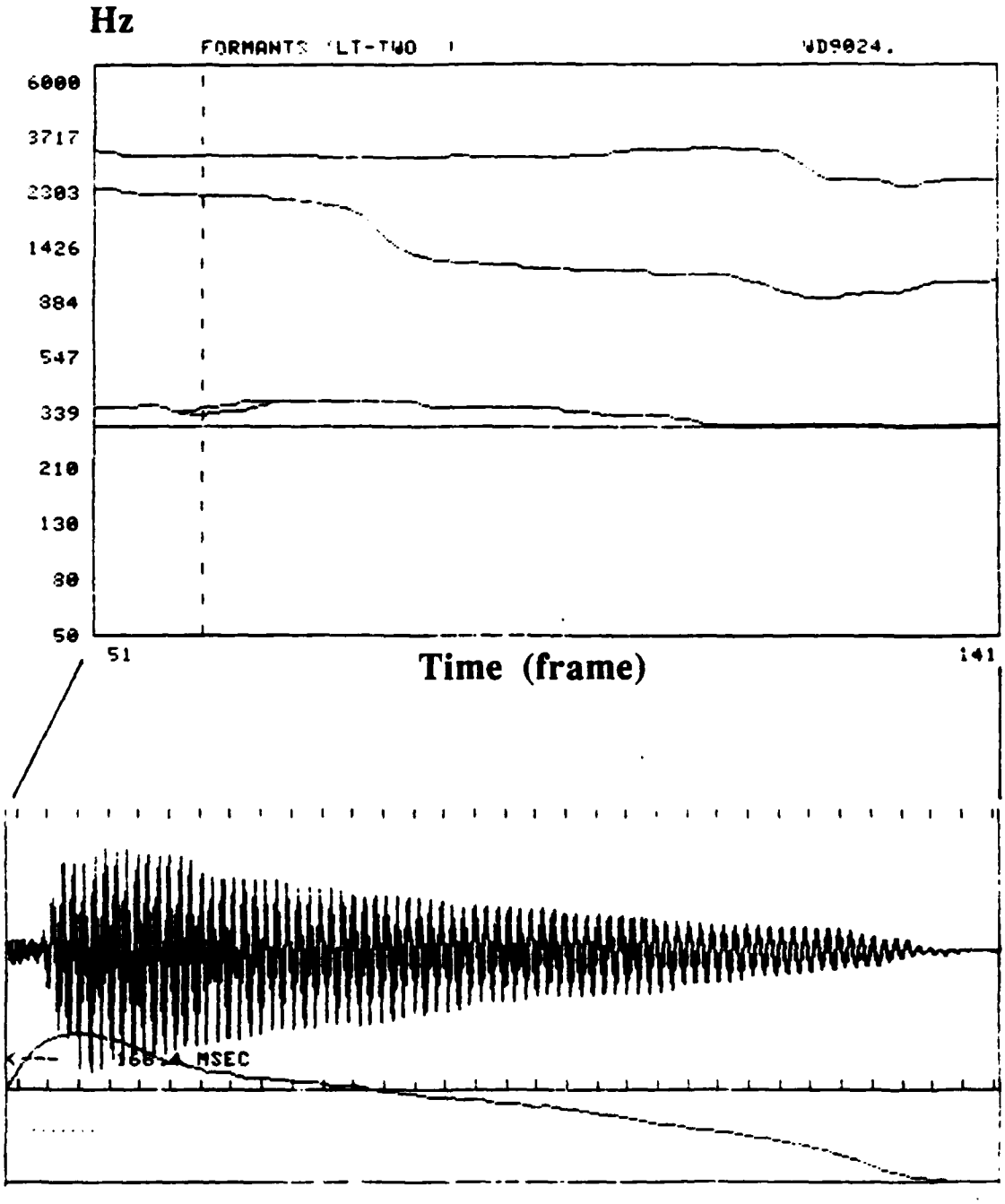
The following plots are made on a Hewlett-Packard 2623A graphical terminal. The sensory formant contours are generated from the strongest glottal-source (SGS) segments of the utterances. The waveforms corresponding to each SGS are also plotted under the sensory formant contours. Note that under each waveform plot, there are two curves plotted along the time domain. The dotted line represents pitch values while the solid line is the unsmoothed signal amplitude.



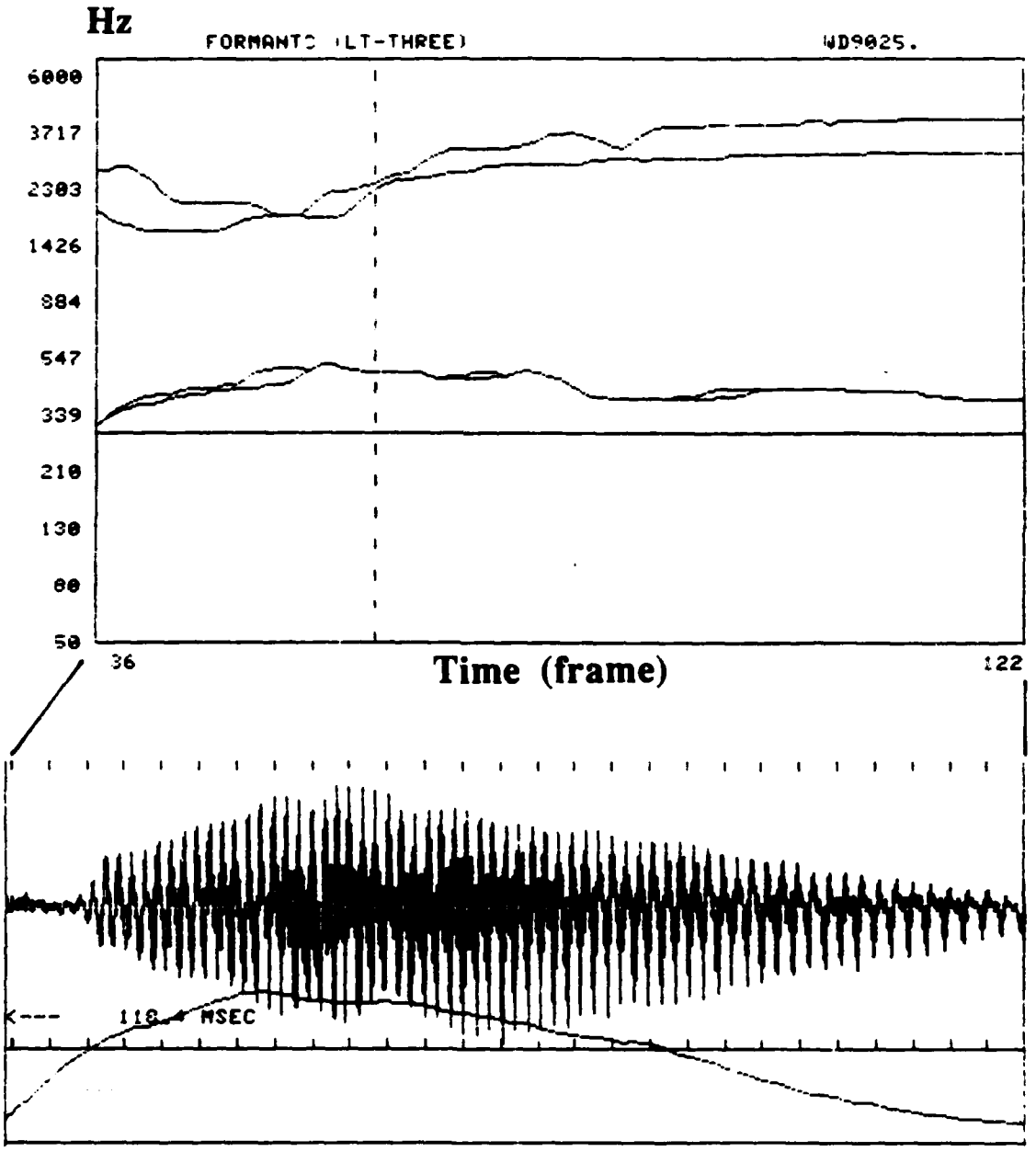
ZERO: female utterance



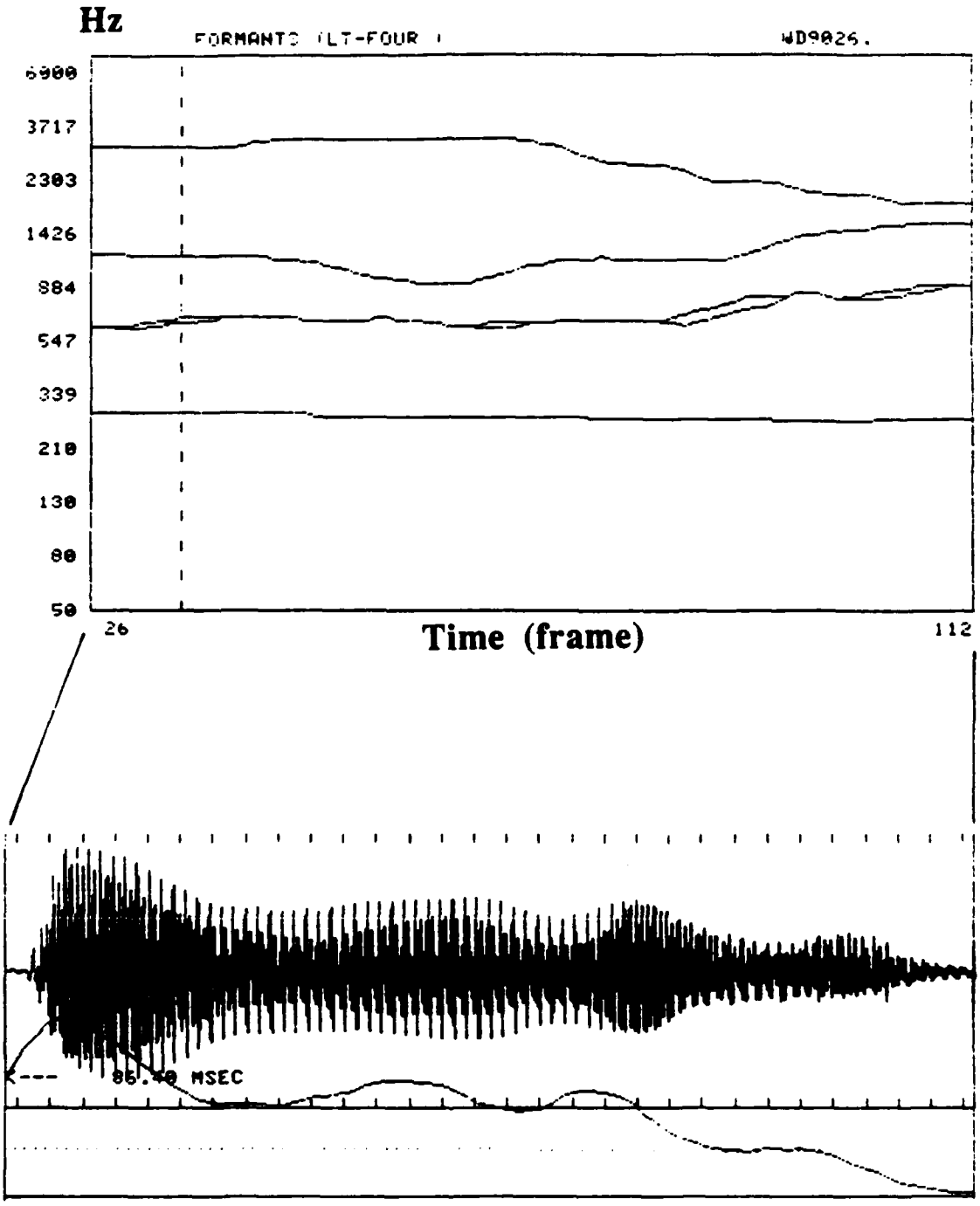
ONE: female utterance



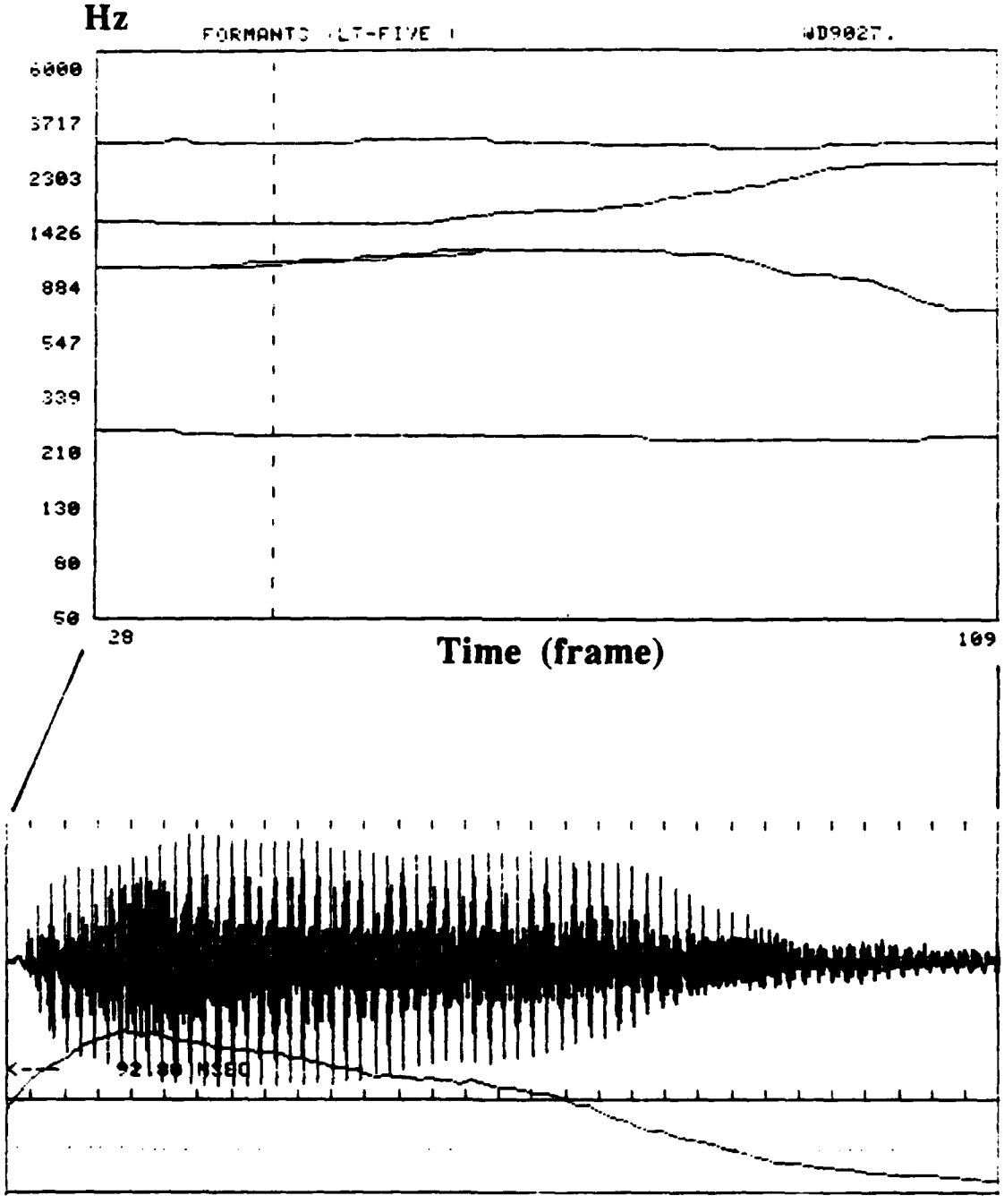
TWO: female utterance



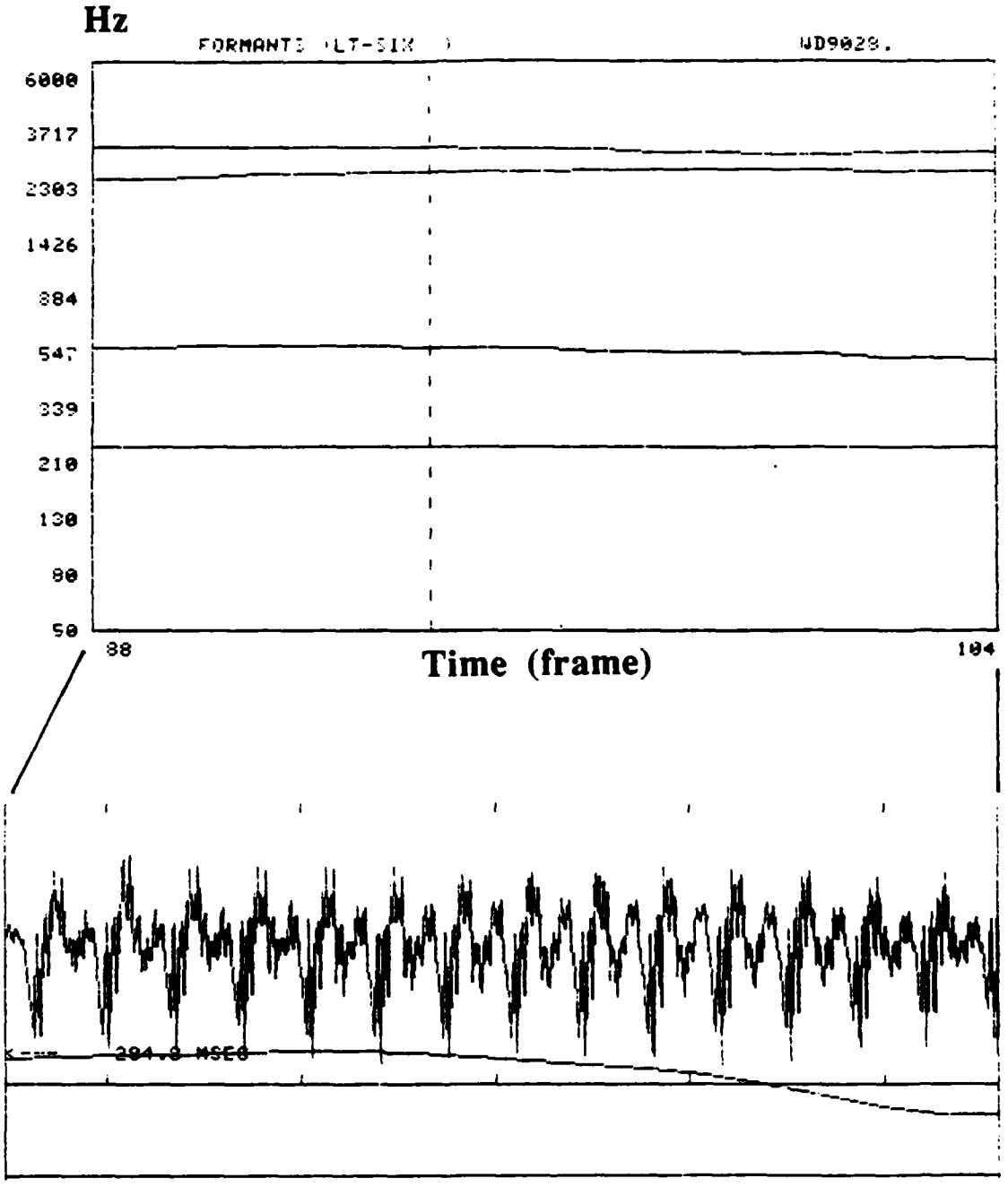
THREE: female utterance



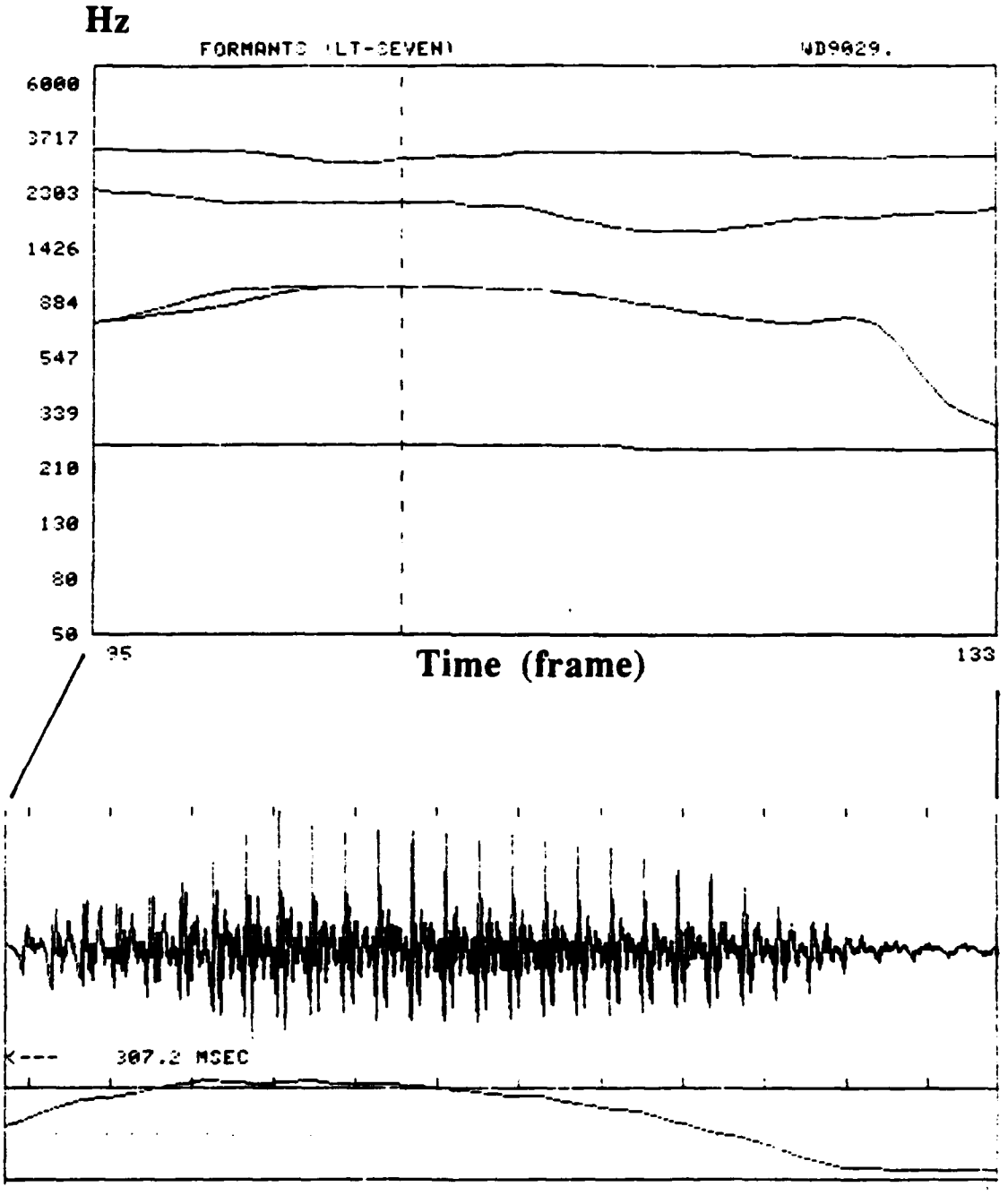
FOUR: female utterance



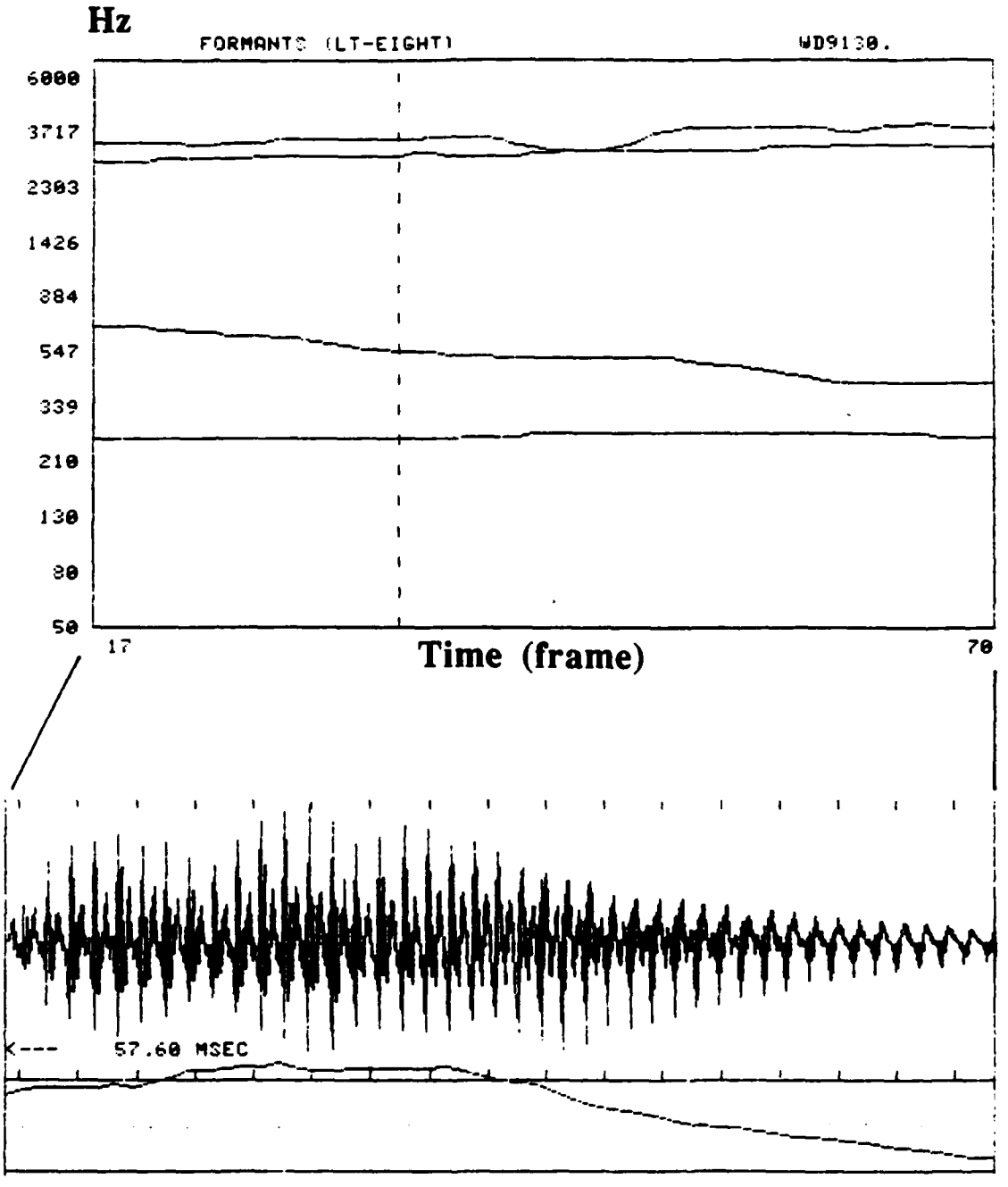
FIVE: female utterance



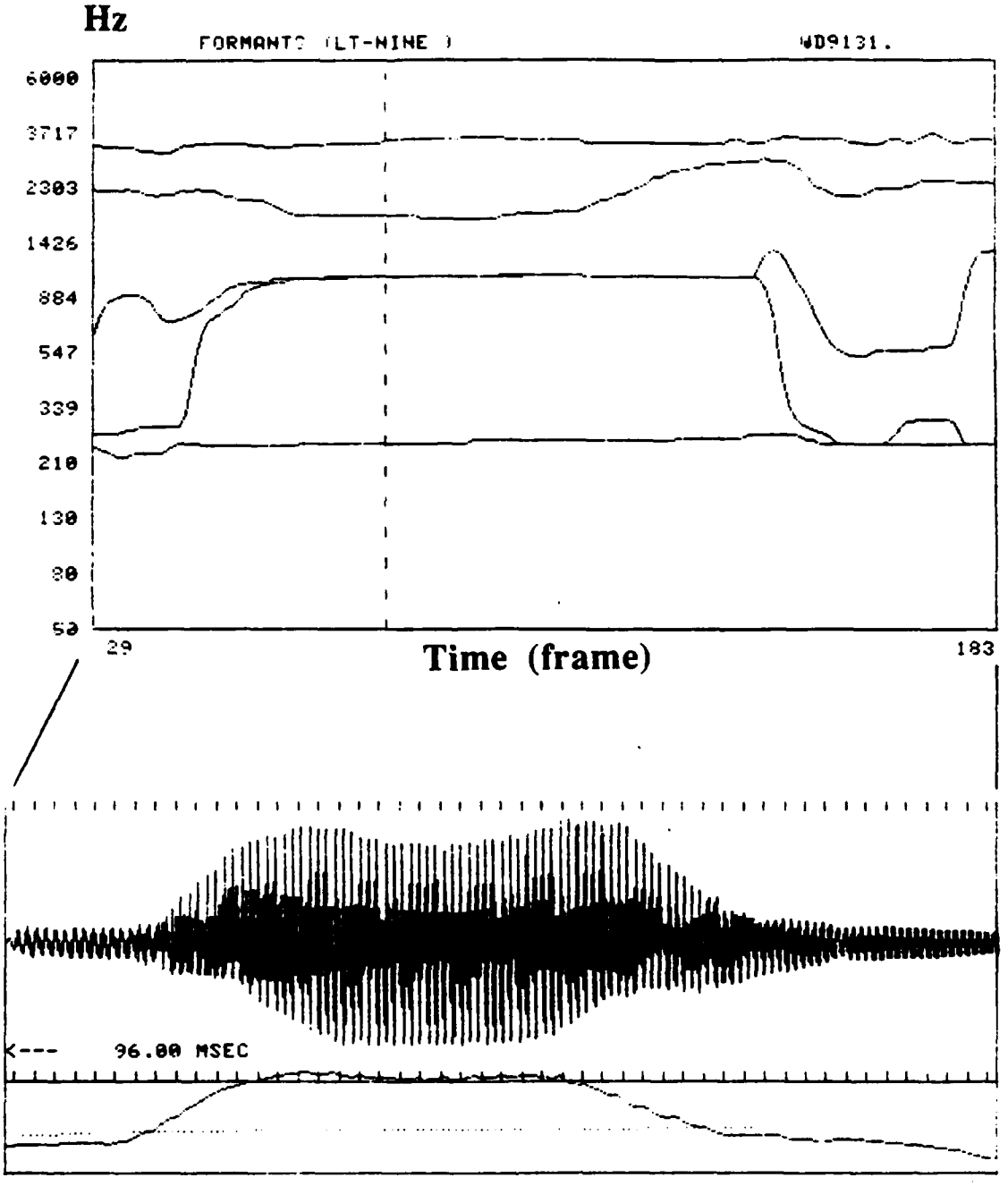
SIX: female utterance



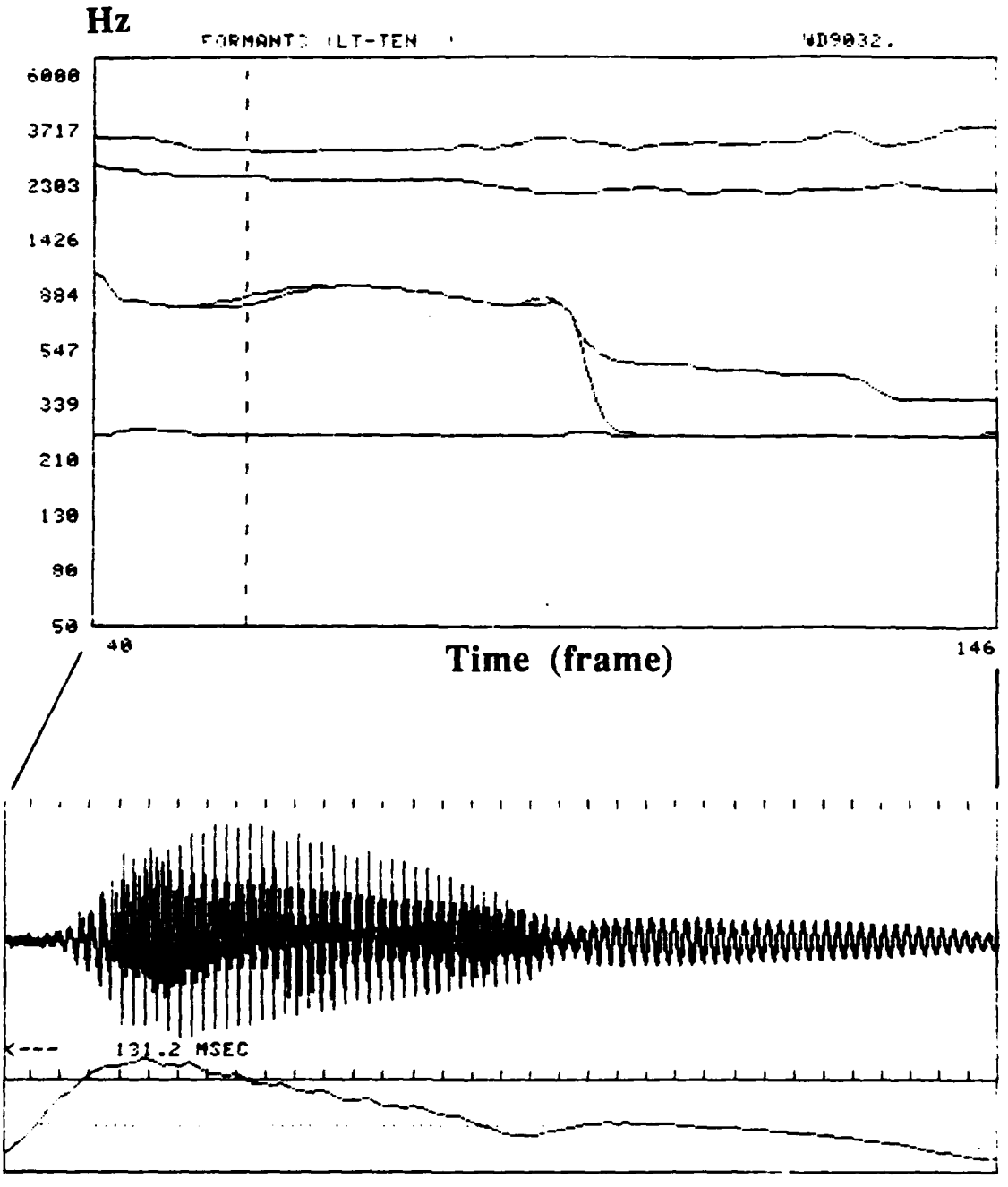
SEVEN: female utterance



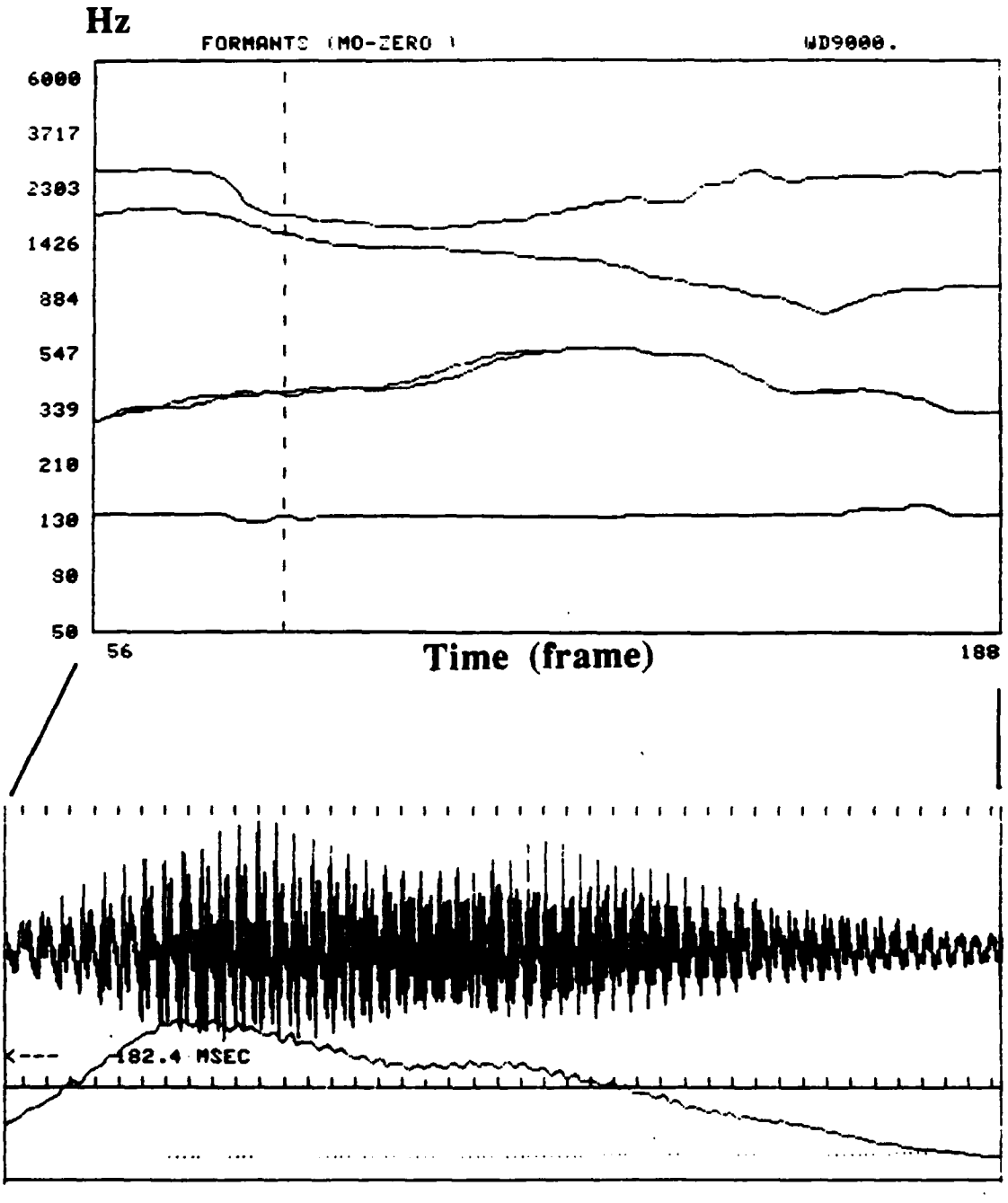
EIGHT: female utterance



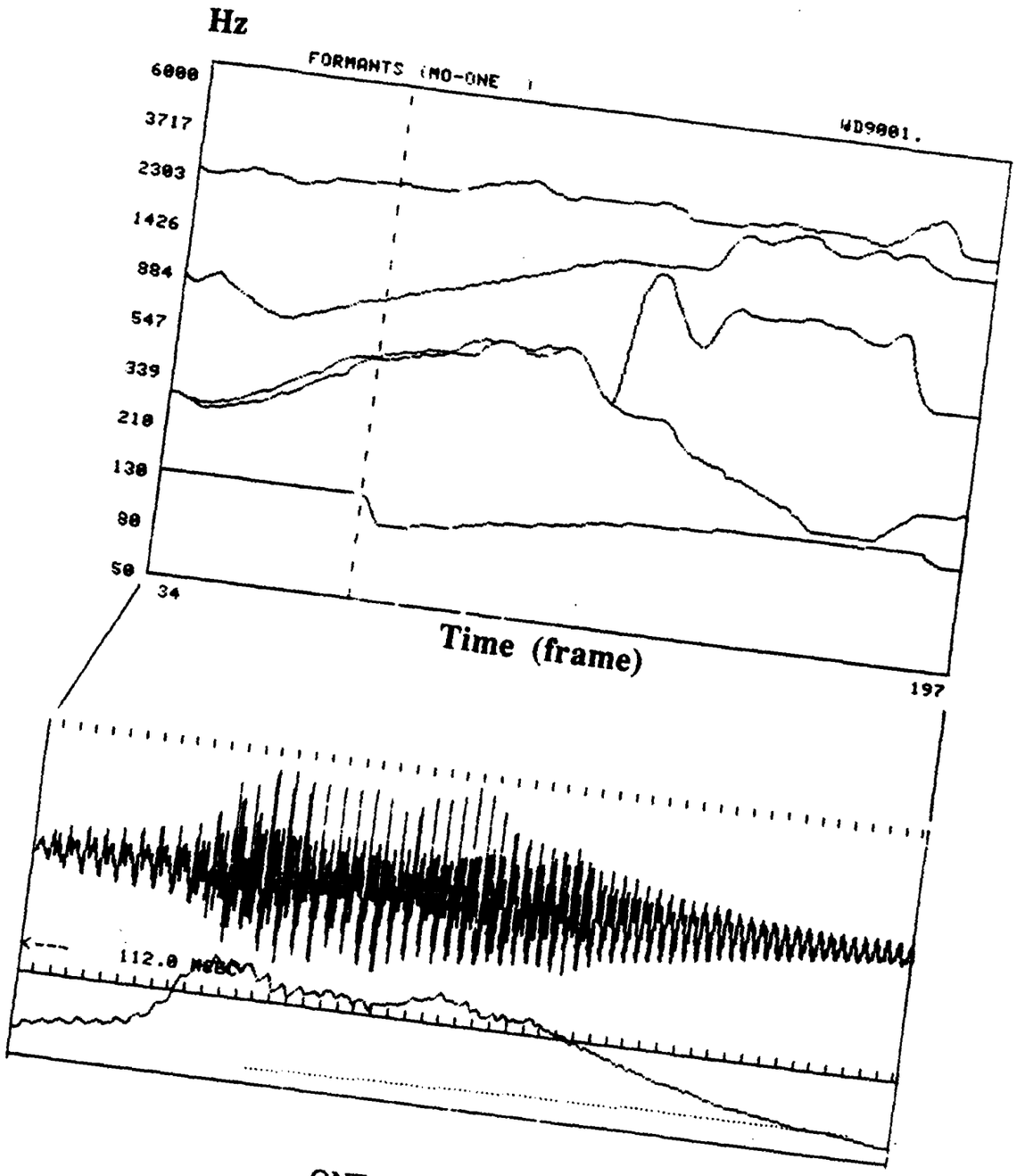
NINE: female utterance



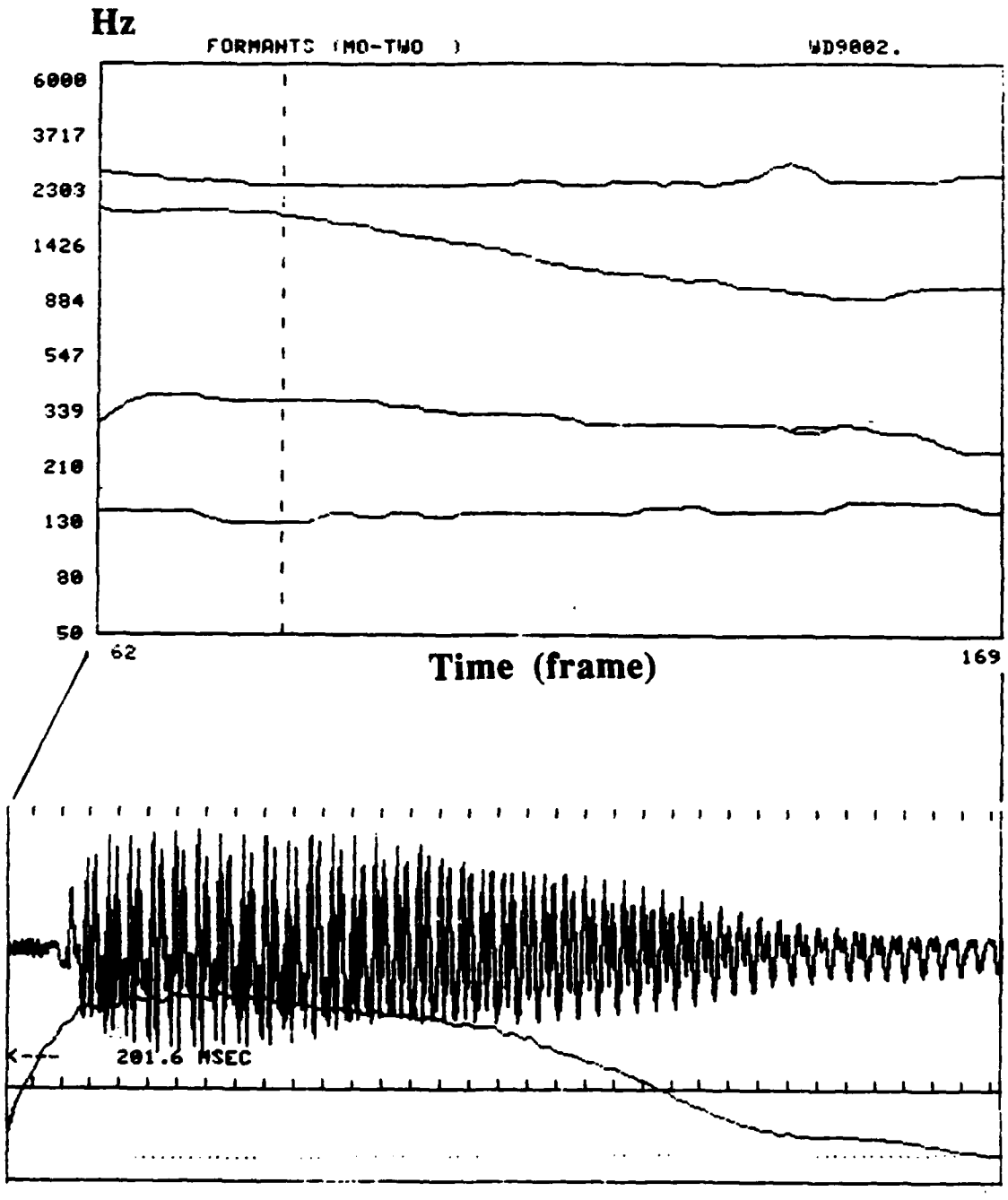
TEN: female utterance



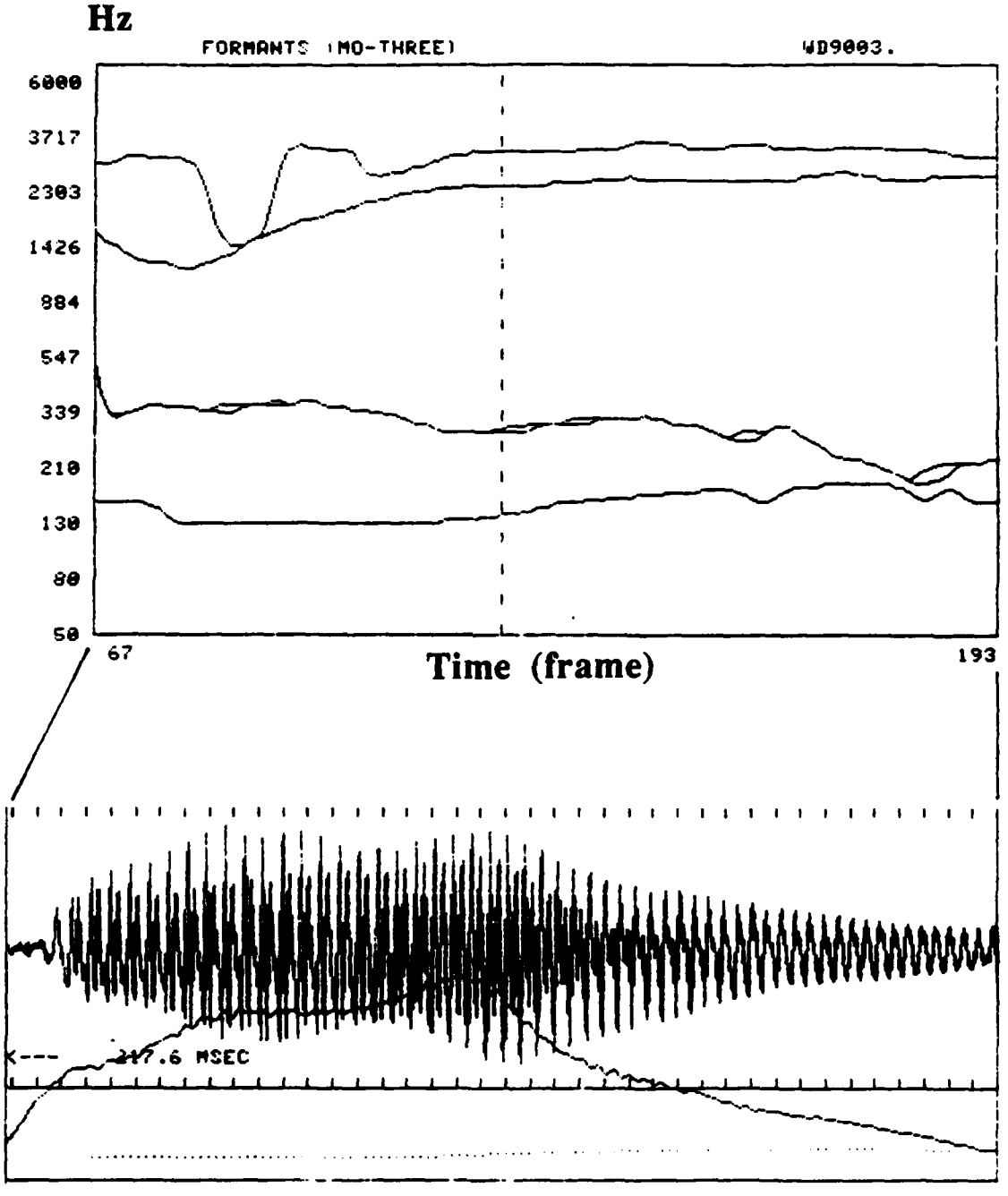
ZERO: male utterance



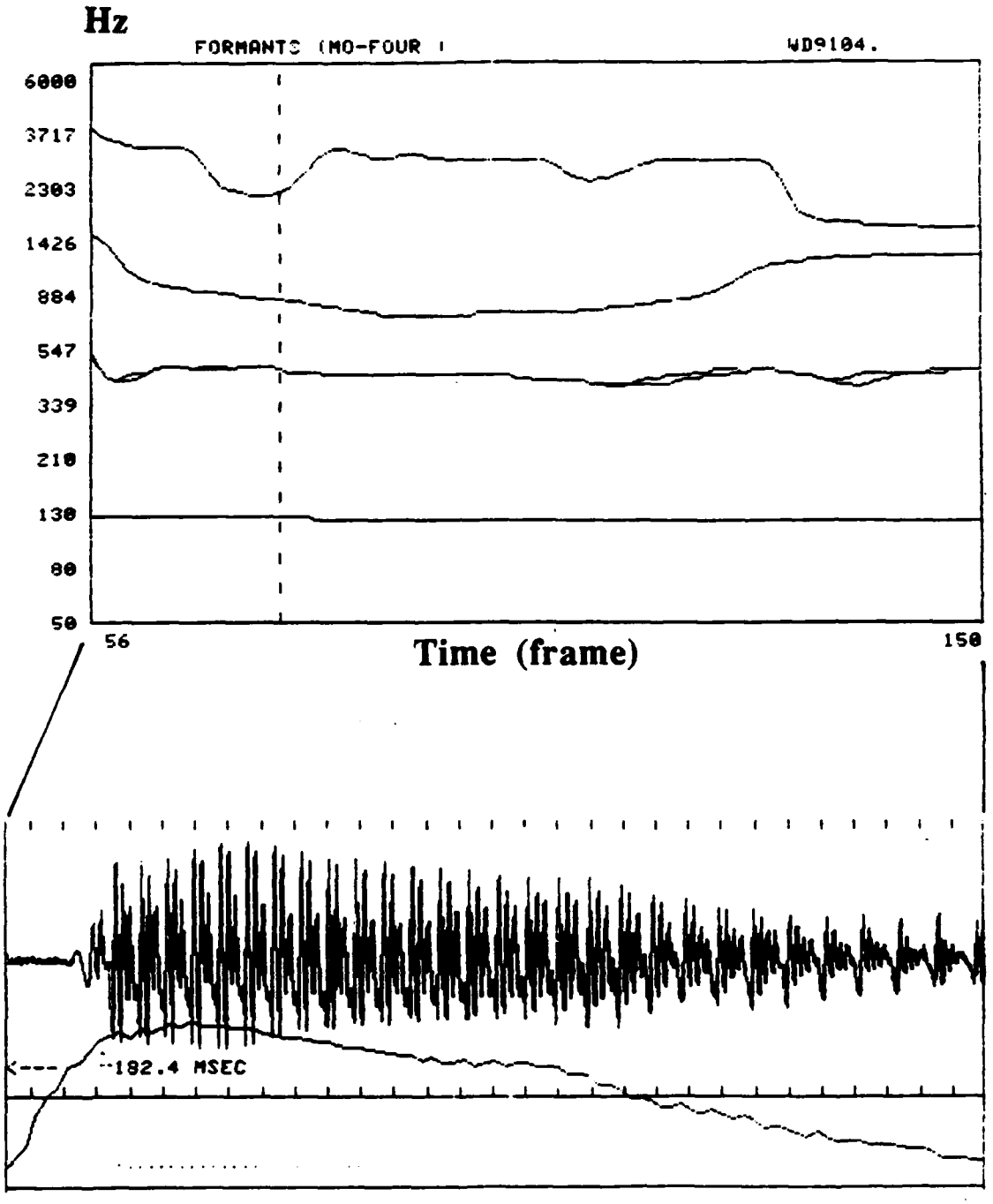
ONE: male utterance



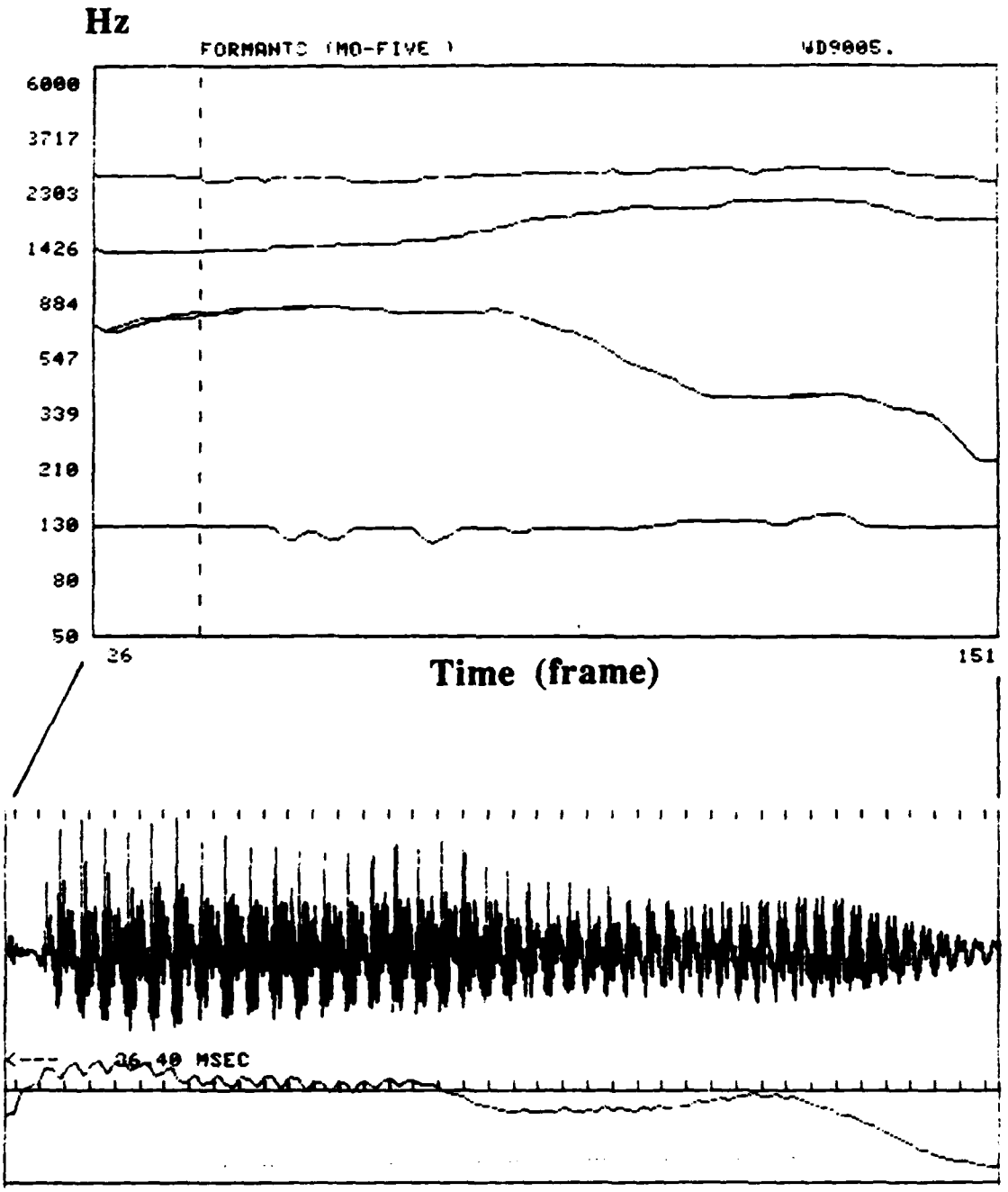
TWO: male utterance



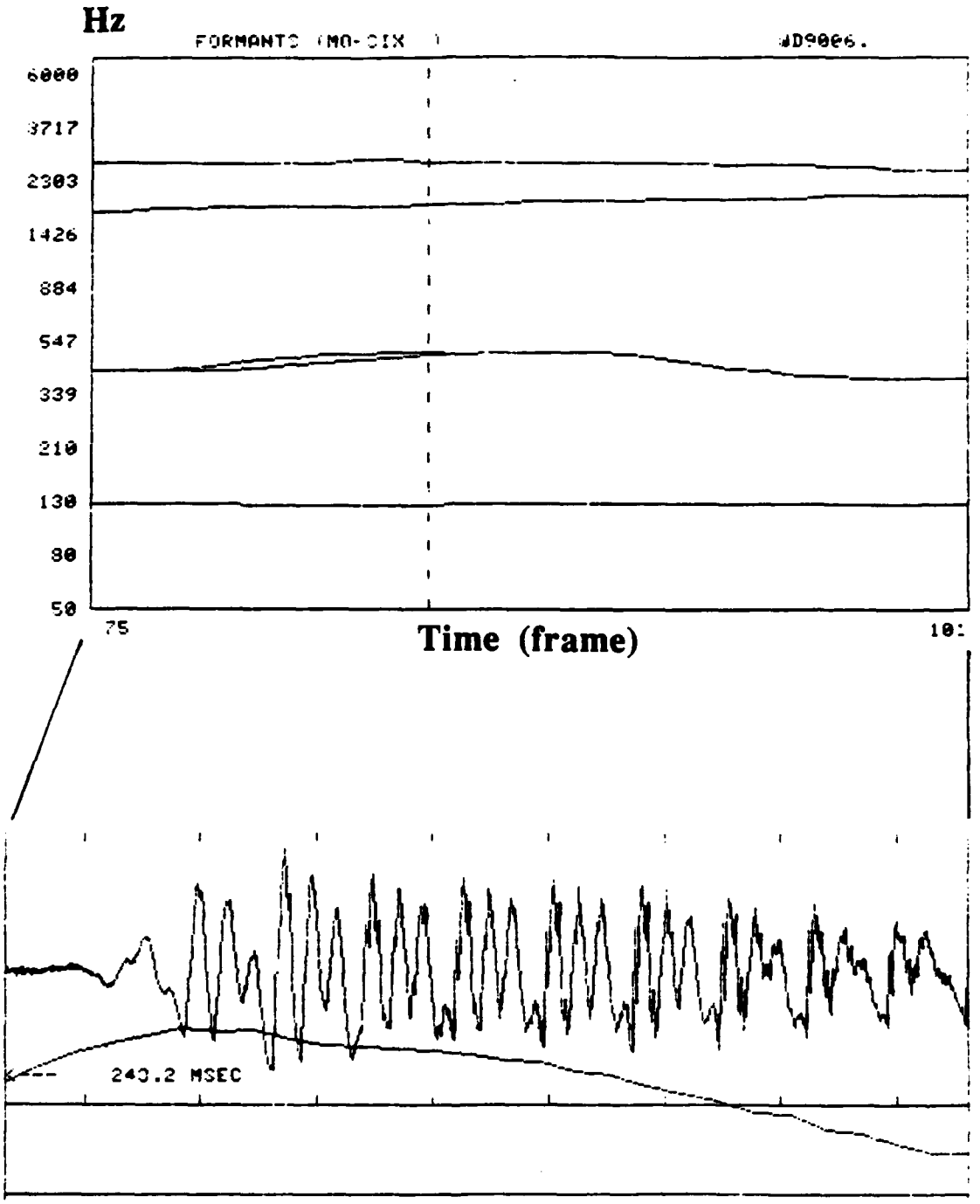
THREE: male utterance



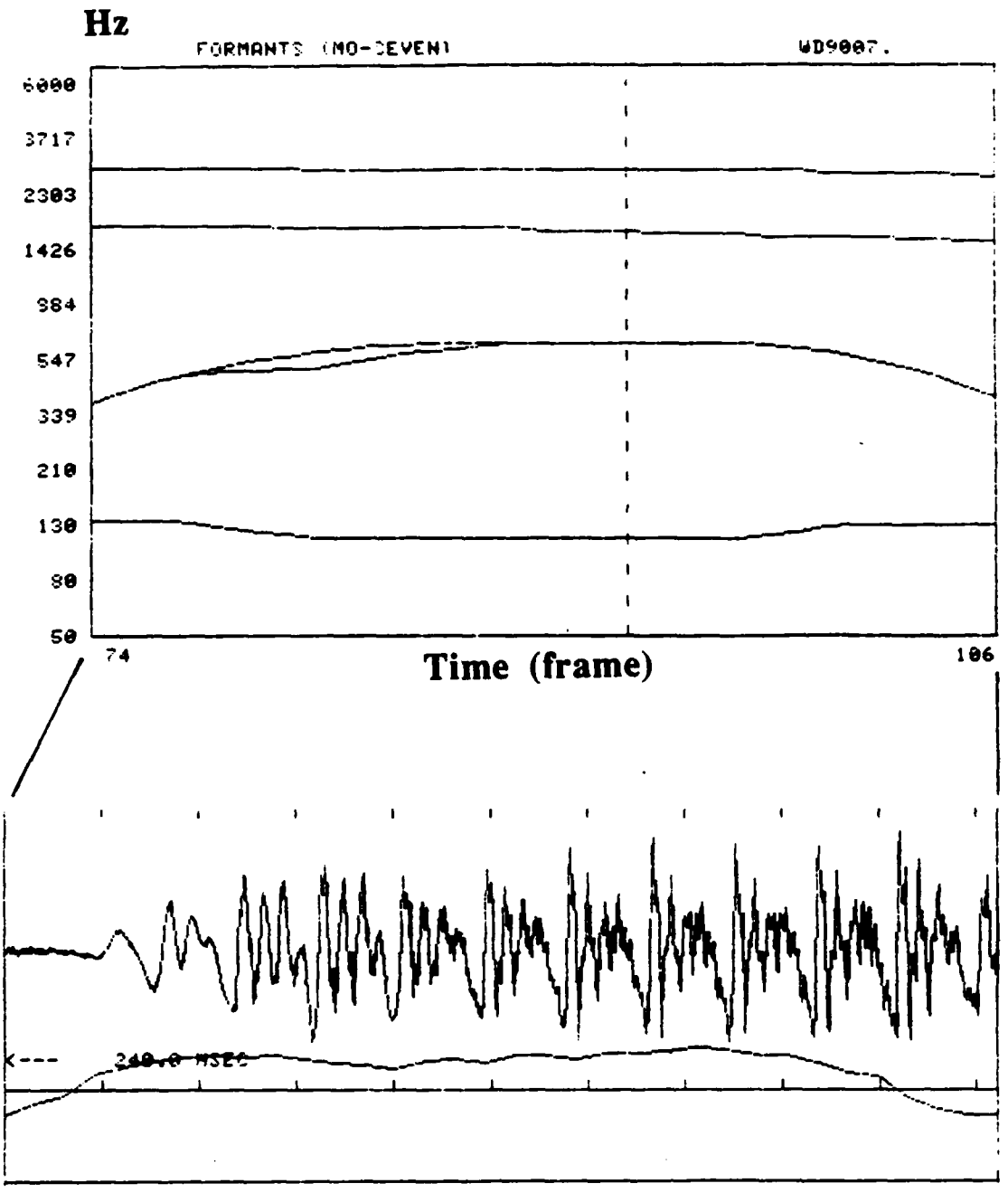
FOUR: male utterance



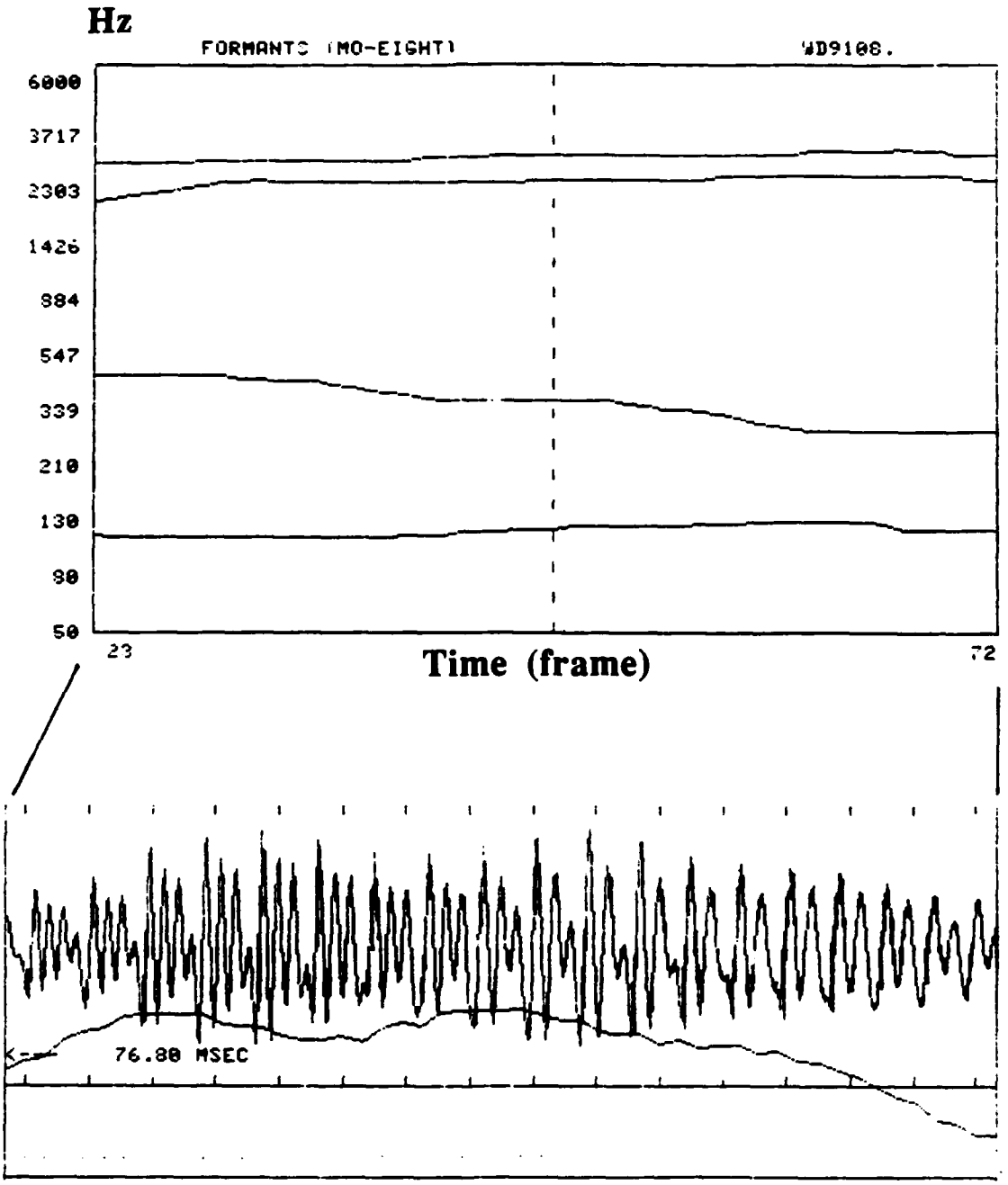
FIVE: male utterance



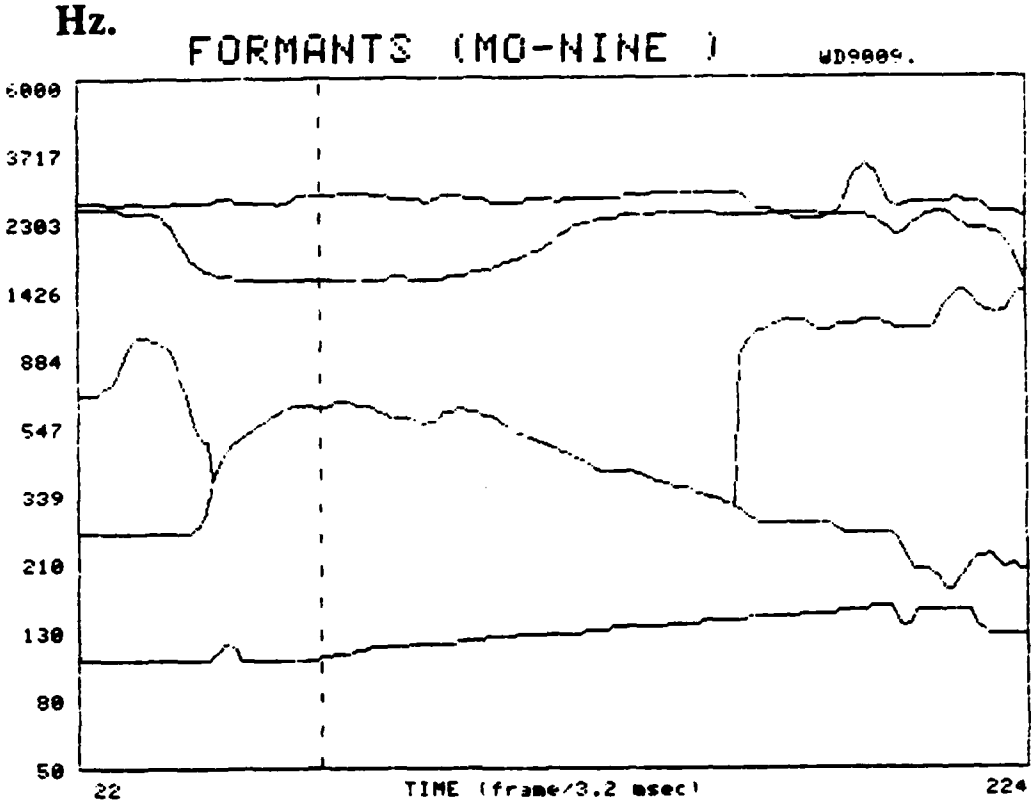
SIX: male utterance

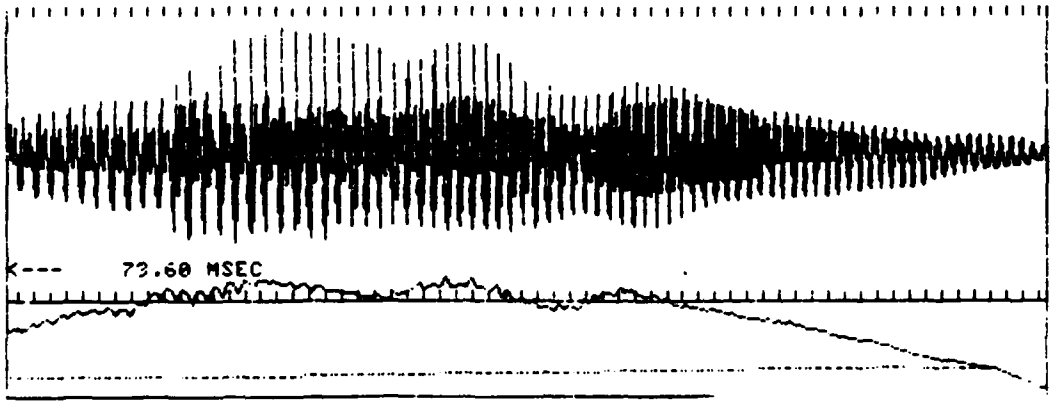


SEVEN: male utterance

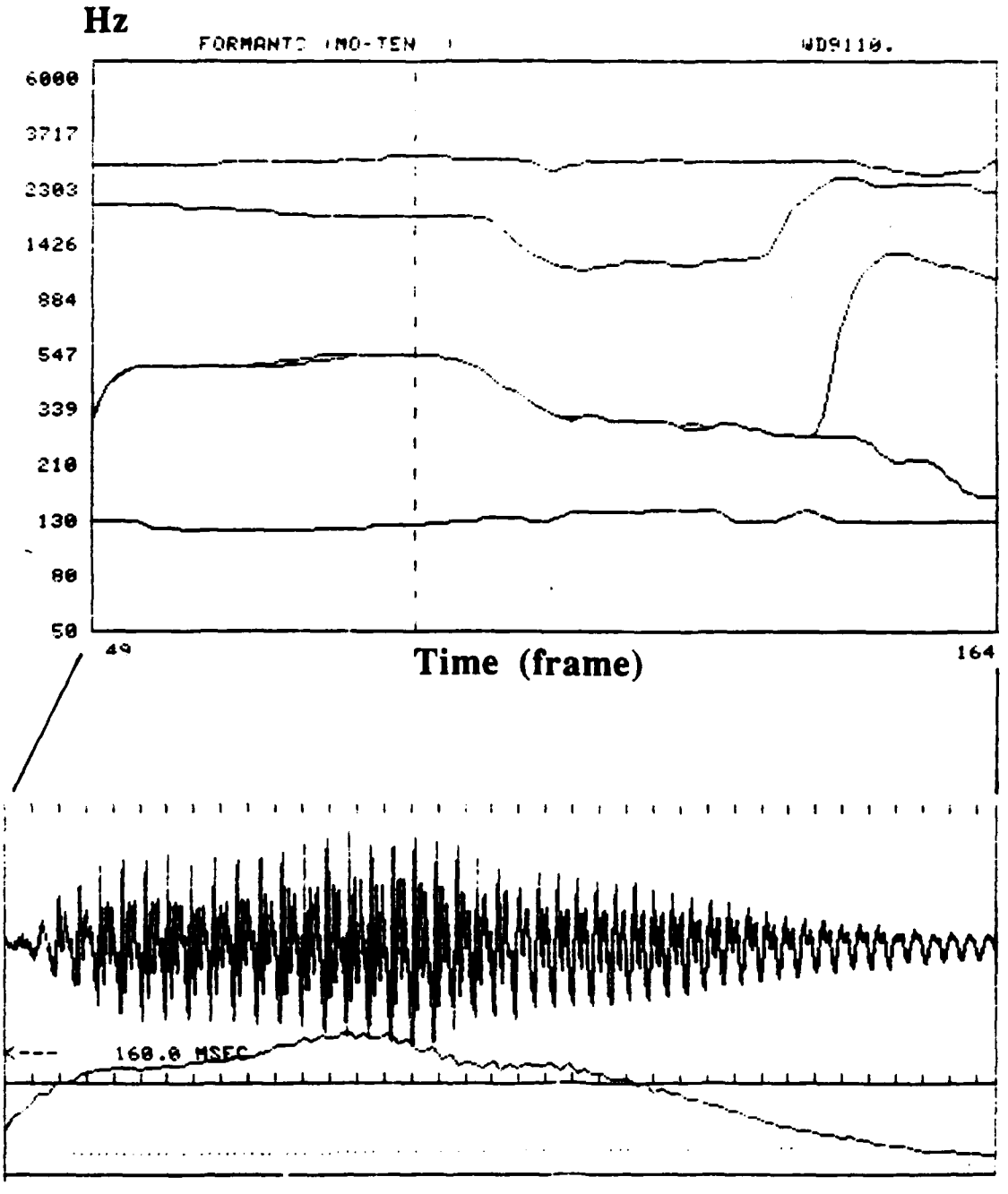


EIGHT: male utterance





NINE: male utterance



TEN: male utterance

11. BIBLIOGRAPHY

1. Lee, W.A. and Shoup, J.E., "Specific Contribution of the ARPA SUR Project," Trends in Speech Recognition, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

2. Walker, D.E., "The SRI Speech Understanding System," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 23, Number 5, October 1975.

3. Woods, W.A., et al., "Speech Understanding System -- Final Technical Progress Report," Report Numer 3438, Volume I-V, Bolt Berbanek and Newman Inc., Cambridge, Massachusetts, 1976.

4. Reddy, D.R., et al., "Speech Understanding System: Summary of Results of the Five Year Research Effort at CMU," pp 1039-1050, 1979.

5. Barnett, J.A., et al., "The SDC Speech Understanding System," Trends in Speech Recognition, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

6. Zue, V.W., "The Use of Phonetic Rules in Automatic Speech Recognition," Speech Communication, Volume 2, Number 2-3, North-Holland, July, 1983.

7. Lee, W.A., "Speech Recognition: What Is Needed Now ?" Trends in Speech Recognition, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

8. Reddy, D.R. and Zue, V.W., "Recognition Continuous Speech Remains an Elusive Goal," IEEE Spectrum, November 1983.

9. Shirai, K. and Honda, M., "Estimation of Aritculary Motion from Spoken Language Generation and Understanding," Proceedings of the NATO Advanced Study Institute, Bonas, France, June 26 - July 7, 1979.

10. Klatt, D.H., "Script and Lafs: Two New Approaches to Speech Analysis," Trends in Speech Recognition, pp 1039-1050, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

11. Rosenberg, A.E., et al., "Demisyllable-Based Isolated Word Recognition System," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 23, Number 5, October 1983.

12. Tanaka Atsto, et al. and Yamashita Kazumi, "A Study of the Syllable Oriented Recognition of Continuous Speech," Speech Communication, Volume 2, Number 2-3, pp 207-210, July 1983.

13. Bridle, J.S. and Chamberlain, R.M., "Automatic Labelling of Speech Using Synethsis-by-Rule and Non-linear Time-Alignment," Speech Communication, Volume 2, Number 2-3, pp 187-189, July 1983.

14. Cole, R.A., et al., "Speech As Patterns on Paper," Perception and Production of Fluent Speech, Lawrence Erlbaum Associates, Publishers. Hillsdale, New Jersey, pp 3-42, 1980.

15. Zue, V.W. and Cole, R.A., "Experiments on Spectrogram Reading," Conference Record, the IEEE International Conference on Acoustics, Speech and Signal Processing, pp 116-119, 1979.

16. Zue, V.W. and Laferriere, M., "Acoustic Study of Medial /t, d/ in American English," Journal of Acoustical Society of America, Volume 66, Number 4, pp 1039-1050, 1979.

17. Zue, V.W. and Shattuck-Hufnagel, S., "Palatalization of /s/ in American English: When is a /s/ not a /s/ ?", Journal of Acoustical Society of America, Volume 67, (S27), 1980.

18. Stefik, M., "Strategic Computing at DARPA : Overview and Accessment," Communication of ACM, Volume 28, Number 7, 1985.

19. Fischetti, M.A., "U.S. Research: Foresting the next generation of computing," IEEE Spectrum, pp 59-62, November 1983.

20. Makhoul, J., Schwartz, R., and et al., "Continuous Speech Recognition," Journal of Acoustical Society of America, Volume 76,, (S-46), October 1984.

21. Aull, A.M. and Zue, V.W., "Lexical Stess and Its Application in Large Vocabulary Speech Recognition", Journal of Acoustic Society of America, Volume 76, (S-47), October 1984.

22. Burton, D.K., Shore, J.E. and Buck, J., "A Generalization of Isolated Word Recognition Using Vector Quantization," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Boston, Massachusetts, April 1983.

23. Miller, J.D., "Auditory Processing of the Acoustic Patterns of Speech", Archives of Otolaryngology, Volume 110, March 1984.

24. Chang, H.M. and Miller, J.D, "Automatic Target Verification for Vowels," Journal of Acoustical Society of America, Volume 79, (S1), Spring 1986.

25. Rudnicky, A.I., "Speaker-Independent Recognition of Vocalic Segments," Journal of Acoustical Society of America, Volume 76, (S1), October 1984.

26. Sakoe, H. and Chiba, S., "A Dynamic Programming Approach to Continuous Speech Recognition," Proceedings of International Congress of Acoustics, Budapest, Hungary, (Rep. 20-C-13), 1971.

27. Sakoe, H. and Chiba, S., "Comparative Study of DP-Pattern Matching Techniques for Speech Recognition," Speech Resource Group, Acoustical Society of Japan, (Rep. S73-22), 1973.

28. Meisel, W.S., "Implication of Large Vocabulary Recognition," Speech Tech 1986, Conference Proceedings, April 28-30, 1986.

29. Lesser, V.R., Fennell, R.D., Erman, L.D. and Reddy, D.R., "Organizations of the HEARSAY-II Speech Understanding System," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 23, Number 1, pp 72-79, February, 1983.

30. Lowerre, B.T., "The Harpy Speech Understanding System," Ph.D Dissertation presented to the Carnegie-Mellon University, Pittsburgh, Pennsylvania, 1976.

31. Bahl, L.R., Jelinek, F. and Mercer, R.L., "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence," PAMI-5:179-190, 1983.

32. Myers, C.S. and Levinson, S.E., "Speaker-Independent Connected Word Recognition Using a Syntax-Directed Dynamic Programming Procedure," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 30, pp 561-565, 1982.

33. Earley, J., "An Efficient Context-Free Parsing ALgorithm," Communication of ACM, Volume 13, pp 94-102, 1970.

34. Nakagawa, S., "A Machine Understanding System for Spoken Japaness Sentences," Ph.D thesis, Kyoto University, 1976.

35. Sakai, T. and Nakagawa, S., "Speech Understanding System LITHAN and Some Applications," Proceedings of the 3rd International Conference on Pattern Recognition, pp 621-625, 1976.

36. Meisel, W.S., personal communication, 1986.

37. Colla, A.M., Scagliola, C. and Sciarra, D., "A Connected Speech Recognition System Using a Diphone-Based Language Model," Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp 1229-1232, 1985.

38. Flanagan, J.L., "They are coming, Computer that hear and response human speech," Record, AT&T Bell Laboratories, Murry Hills, New Jersey, (3-3), November 1983.

39. Reddy, D.R., "Speech Recognition by Machine: A Review," Proceedings of IEEE, Volume 64, pp 501-531, 1976.

40. Nye, J.M., "Expanding Market for Commercial Speech Recognition," Trends in Speech Recognition, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

41. Fujimura, O., "Syllable as Unit of Speech Recognition," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 23, Number 1, pp 82-87, February, 1975.

42. Peterson, G.E. and Lehiste, I., "Duration of Syllable in English," Journal of Acostical Society of America, Volume 32, Number 6, June 1960.

43. Massaro, D.W., "Preperceptual Images, Processing Time, and Perceptual Units in Auditory Perception," Psychological Review, Volume 79, pp 123-145, 1978.

44. Huggins, A.W.E., "Distortion of the Temporal Pattern of Speech: Interruption and Alternation," Journal of Acostical Society of America, Volume 36, pp 1055-1064, 1960.

45. Studdert-Kennedy, M., "Speech Perception", Contemporary Issues in Experimental Phonetics, Academic Press, New York, 1976.

46. Rosenberg, A.E., Rabiner, L.R., Levinson, S.E., and Wilpon, J.G., "A Preliminary Study on the Use of Demisyllables in Automatic Speech Recognition," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp 967-970, Atlanta, Georgia, 1981.

47. Ruske, G. and Schotola, T., "The Efficiency of Demisyllable Segmentation in Recognition of Spoken Words," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp 971-974, Atlanta, Georgia, 1981.

48. Denes, P.B. and Pinson, E.N., The Speech Chain: Physics and Biology of Spoken Language, Anchor Press/Doubleday, Garden City, New York, pp 14-14, 1973.

49. Liberman, A.M., Cooper, F.S., Shankweiler, D.P. and Studdert-Kennedy, M., "Perception of Speech Code," Psychological Review, Volume 74, Number 6, pp 431-460, November 1967.

50. Gallanher, R.T., "Phonemic Strategy Guides the Search for Multilingual Speech Recogntion Systems," Electronics, May 19, 1983.

51. Klatt, D.H., "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access," Perception and Production of Fluent Speech, (Cole ed.), Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp 243-287, 1980.

52. Leung, H.C. and Zue, V.W., "A Procedure for Automatic Alignment of Phonetic Transcription with Continuous Speech," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (2.7), San Diego, California, March 1984.

53. Shore, J.E. and Burton, D.K., "Discrete Utterance Recognition without Time Alignment," IEEE Transactions on Information Theory, Volume 4, pp 473-491, July 1983.

54. Bush, M.A., Kopec, G.E. and Lauritzen, N., "Segmentation in Isolated Word Recognition Using Vector Quantization," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (17.11), San Diego, California, March 1984.

55. Chomsky, N. and Halle, M., The Sound Pattern of English, Harper & Row Publishers, Inc., Hagerstown, Maryland, 1968.

56. Stevens, K.N., "Study of Acoustic Properties of Spech Sound II and Some Remarks on the Use of Acoustic Data in Schema for Machine Recognition of Speech," Scientific Report Report Number 12, Report AFCR L-69-0039, Bolt Berankek and Newman, Cambridge, Massachusetts, 1969.

57. Flanagan, J.L., Speech Analysis, Synthesis, and Perception, Springer, New York, 1972.

58. Rabiner, L.R. and Gold, B., Theory and Application of Digital Signal Processing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, pp 26-28, 1975.

59. Fant, G., Speech Sounds and Features, pp 277, MIT press, Cambridge, Massachusetts, 1973.

60. Lasry, M.J. and Stern, R.M., "Unsupervised Adaptation to New Speakers in Feature-based Letter Recognition," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (17.6.1-17.6.4), San Diego, California, March 1984.

61. Cole, R.A., et al., "Feature-based Speaker-Independent Recognition of Isolated English Letters," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1983.

62. Cole, R.A., Stern, R.M. and Lasry, M.J., "Performing Fine Phonetic Distinctions: Templates vs. Features," Invariance and Variability of Features in Spoken English Letters, (J. Perkell, et al., ed.), Lawrence Erlbaum, New York, 1983.

63. Miller, N.J., "Pitch Detection by Data Reduction," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 23, Number 1, pp 72-79. February, 1975.

64. Rabiner, L.R., Cheng, M.J., Rosenberg, A.E. and McGonegal, C.A., "Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 24, pp 399-417, 1976.

65. Seneff, S.S., "Real Time Harmonic Pitch Detector," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 2, (2.7), San Diego, California, March 1984.

66. Specker, P., "A Powerful Post-Processing Algorithm for Time-Domain Pitch Tracker," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (2.7), San Diego, California, March 1984.

67. ~~Michael, S.~~ Phillips, M.S., "A Feature-based Time Domain Pitch Tracker," Journal of Acoustical Society of America, Volume 77, (S1), Spring 1985, p. S9,

68. Markel, J.D., "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation," IEEE Transactions on Audio Electroacoustics, Volume AU-20, pp 129-137, June 1972.

69. McCandless, S.S., "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Transactions on Acoustics, Speech and Signal Processing, April 1974.

70. Markel, J.D. and Gray, A.H. Jr., Linear Prediction of Speech, Springer, New York, 1976.

71. Reddy, N.S. and Swamy, M.N.S., "High-Resolution Formant Extraction from Linear-Prediction Phase Spectra," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 32, Number 6, pp 1136-1144, December 1974.

72. Shozo, Markino and Ker'ini Kido, "A Speaker-Independent Word Recognition System Based on Phoneme Recognition for a Large Size (212 Words) Vocabulary," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (17.7), San Diego, California, March 1984.

73. Fonsale Patrik, "Feature-based Speaker-Independent Word Recognition without Oral Learning," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (17.7), San Diego, California, March 1984.

74. Kopec, G.E., "Voiceless Stop Consonant Identification Using LPC Spectra," Fairchild Laboratory for Artificial Intelligence Research, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (17.7), San Diego, California, March 1984.

75. Kumar, A. and Bekey, G.A., "Recognition of Consonants Using ARMA Model of the Speech Signal," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (17.7), San Diego, California, March 1984.

76. Davis, H.K., Biddulph, R. and Balashek, J., "Automatic Recognition of Spoken Digits," Journal of the Acoustical Society of America, Volume 24, pp 637-645, 1952.

77. Weinstein, C.J., McCandless, S.S., Mondshein, L.F. and Zue, V.W., "A System for Acoustic-Phonetic Analysis of Continuous Speech," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 23, pp 54-67, 1975.

78. Woods, W.A. and Zue, V.W., "Dictionary Expansion via Phonological Rules for a Speech Understanding Systgem," Teacher, pp 561-564, 1976.

79. Kameny, I., "Automatic Acoustic-Phonetic Analysis of Vowels and Sonorants," Teacher, pp 166-169, 1976.

80. Culter, A. and Foss, D.J., "On the Rule of Sentence Stress in Sentence Processing," Language and Speech, Volume 20, pp 1-10, 1977.

81. Huttenlocher, D.P. and Zue, V.W., "A Model of Lexical Access Based on Partial Phonetic Information," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, (26.4), San Diego, California, March 1984.

82. Anon, "ILS User Guide : V5.0", Signal Technology, Inc. : California, 1985.

83. Flanagan, J.L., "Automatic Extraction of Formant Frequencies from Continuous Speech," Journal of the Acoustical Society of America, Volume 28, pp 110-118, January 1956. (This paperis based on material in "A speech Analyzer for a formant-coding compression system," Sc.D. thesis, MIT, 1955)

84. Pinson, Eilliot, "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths,"Journal of the Acoustical Society of America, Volume 35, pp 1264-1273, August, 1963.

85. Schafer, R.W. and Rabiner, L.R., "System for Automatic Formant Analysis of Voiced Speech," Journal of Acoustical Society of America, Volume 47, pp 637-648, 1970.

86. Chritensen, R.L., Strong, W.J. and Palmer, E.P., "A Comparision of Three Methods of Extracting Resonance Information from Prediction-Coefficients Coded Speech," IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 24, pp 8-14, February 1976.

87. Olive, J.P., "Automatic Formant Tracking by Newton-Raphson Techniques," Journal of Acoustical Society of America, Volume 50, pp 661-670, 1971.

88. Yegnanarayana, B., "Formant Extraction from Linear-Predictive Phase Spectra," Journal of Acoustical Society of America, Volume 63, pp 1638-1642, May 1978.

89. Klatt, D.H., "Software for a Cascade/Parallel Formant Synthesizer." Journal of Acoustical Society of America, Volume 67, pp 971-995, 1980.

90. Miller, J.D. and Chang, H.M., "Sensory-Perceptual Dynamics in Speech Perception," NATO Conference, Utrecht, Netherlands, July 1986.

91. Miller, J.D., "Auditory-Perceptual Correlates of the Vowels." Journal of Acoustical Society of America, Volume 76(S1), pp S79., 1984.

92. Miller, J.D. and Hawks, J., "Auditory-Perception of English Consonants." Journal of Acoustical Society of Amerca, Volume 79(S1), 1986.

93. Potter, R.K., and Kopp, G.A., and Green, Harriet, C., Visible Speech. New York, Van Nostrand, 1947.

94. Potter, R.K. and Peterson, G.E., "The Representation of Vowels and Their Movements." Journal of Acoustical Soceity of Amerca, Volume 20, pp 528-535, 1948.

95. Joos, M., "Acoustic Phonetics." Language Monogr, Volume 23, 1948.

96. Liddell, M. H., "The Physical Characteristics of Speech Sound." Bull. No. 16, Eng. Exper. Stat., Purdue University, 1924.

97. Liddell, M.H., "The Physical Characteristics of Speech Sound-II." Bull. No. 23, Eng. Exper. Stat., Purdue University, 1925.

98. Liddell, M.H., "The Physical Characteristics of Speech Sound-III." Bull. No. 28, Eng. Exper. Stat., Purdue University, 1927.

99. Holbrook, A. and Fairbanks, G., "Diphthong Formants and Their Movements." Journal of Speech and Hearing Research, March, pp 249-269, 1962.

100. Cole, R.A., personal communication, December 1986.

101. Miller, J.D., "Auditory Processing of the Acoustic Patterns of Speech." Arch. Otolaryngol. Volume 110, pp 154-159, 1984.

102. Fletcher, H., "Auditory Patterns." Review of Modern Physics, Volume 12, pp 47-65, 1940.

103. Munson, W.A and Gardner, M.B., "Loudness Patterns - a New Approach." Journal of Acoustical Soceity of Amerca, Volume 22, pp 177-190, 1950.

104. Plomp, R., "Timbre as a Multidimensional attribute of Complex Tones," Frequency Analysis and Periodicity Detection in Hearing. Edited by R. Plomp & G.F. Smoorenburg. (A.W. Sijthoff, Leiden), pp 397-414, 1970.

105. Schroeder, M.R., "Modeles of Hearing, " Proceeding of IEEE, Volume 63, pp 1332-1350, 1975.

106. Zwicher, E. and Scharf, B., "A Model of Loudness Summation," Psychological Review, Volume 72 (1), pp 3-26, 1965.

107. Zwichker, E., "Masking and Psychological Excitation as Consequences of the Ear's Frequency Analysis," Frequency Analysis and Preiodicity Detection in Hearing. Edited by R. Plomp & G.F. Smoorenburg. (A.W. Sijthoff, Leiden), pp 376-394, 1970.

108. Studdert-Kennedy, M., Liberman, A.M., and Harris, K.S, "Motor Theory of Speech Perception: A Reply to Lane's Critical Review," Psychological Review, Volume 77, pp 234-249, 1970.

109. Liberman, A.M., "On Finding That Speech Is Special," American Psychology, Volume 38, pp 148-167, 1982.

110. Kozhevnikov, V.A., and Chistovich, L.A., Speech: Articulation and Perception. [Rech: Artikulyatsiya i Vospriyatiye, Moscow-Lenigrad.] Translated by the Joint Publications Research Service. Clearinghousc for Federal Scientific and Technical Information, U.S. Department of Commerce, Washington, D.C., 20043, (Publication nos. JPRS: 30,543; TT: 65-31233.) Cited in : The Sounds of Speech Communication by J.M. Pickett, University Park Press, Maryland, 1980.

111. Fant, G., Speech Sounds and Features. (Cambridge, MIT Press), pp 277, 1973.

112. Stevens, K.N. and Blumstein, S.E., "The Search for Invariant Acoustic Correlates of Phonetic Features." Perspectives on the Study of Speech, P.D. Eimas and J.L. Miller (Eds.), Erlbaum, Hillsdale, New Jersey, pp 1-38, 1981.

113. Pisoni, D.B. and Sawuusch, J.R., "Some Stages of Processing in Speech Perception." Structure and Process in Speech Perception, A. Cohen and S.G. Nootebaum (Eds.), Springer-Verlag, New York, pp 16-35, 1975.

114. Miller, J.D., "A Phonetically Relevant Auditory-Perceptual Space." Peridic Progress Report No. 25, Central Institute for the Deaf, St. Louis, Missouri, pp 25-26, 1982.

115. Miller, J.D., "Auditory-Perceptual Approaches to Phonetic Perception." Journal of Acoustical Society of America, Volume 71 (S1), pp S112(A), 1982.

116. Miller, J.D., "Characteristics of Vowels Spoken in CVC Contexts," Peridic Progress Report No. 25, Central Institute for the Deaf, St. Louis, Missouri, pp 24-25, 1982.

117. Peterson, G.E., "The Information Bearing Elementes of Speech." Journal of Acoustical Society of America, Volume 24, pp 629-637, 1952.

118. Shepard, R.N., "Psychological Representation of Speech Sounds." Human Communication: A United View, E.E David and P. Denis, Eds., McGraw Hill, New York, pp 67-113, 1972.

119. Pols, L.C.W., Spectral Analysis and Identification of Dutch Vowels in Monosyllablic Words. Institute for Perception TND, Soesterberg, the Netherlands, 1977.

120. Scheffers, M.T.M., "Simulation of Auditory Analysis of Pitch: An Elaboration of the DWS Pitch Meter." Journal of Acoustical Society of America, Volume 74, pp 1716-1725, 1983.

121. Klatt, D.H., "Prediction of Perceived Distance from Critical-band Spectra: A First Step." Proceeding of International Congress of Acoustics, Speech and Signal Processing, Paris, pp 1278-1281, 1982.

123. Miller, J.D., "Implications of the Auditory-Perceptual Theory of Phonetic Perception for Speech Recognition by the Hearing Impaired." ASHA Report, No. 14, pp 45-48, 1984.

124. Marler, P., and Peters, S., "Bird Song and Speech: Evidence for Special Processing," Perspectives on the Study of Speech, edited by P.D. Eimas and J.L. Miller, Erlbaum, Hillsdale, New Jersey, pp 75-112, 1981.

125. Cooper, W.E, Speech Perception and Production, Ablex Publishers Corp., U.S.A., pp 34,94, 1979.

126. Diehl, R.L., "Feature Detectors for Speech: A Critical Reappraisal," Pshchological Bulletin, Volume 89(1), pp 1-18, 1981.

127. Miller, J.D., Wier, C.C., Pastore, R.E., Kelly, W.J., and Dooling, R.J., "Discrimination and Labeling of Noise-buzz Sequences with Varying Noise-lead Times: An Example of Categrical Perception." Journal of Acoustical Society of America, Volume 60, pp 410-417, 1976.

128. Peterson, G.E. and Barney, H.L., "Control Methods Used in the Study of Vowels," Journal of Acoustical Society of America, Volume 24, pp 410-417, 1952.

129. Peterson, G.E., "Parameters of Vowel Quality," Journal of Speech Hearing Research, Volume 4, pp 10-29, 1961.

130. Miller, J.D., "Sensory-Perceptual Dynamics and Categorization in Phonetic Perception." Abstracts of the Sixth Midwinter Meeting of ARO, pp 76, 1983.

131. Lehiste, I., Acoustical Characteristics and Selected English Consonants. International Journal of American Linguistics, Volume 30, 1964.

132. Fourakis, M., Hawks, J., and Miller, J.D., "Speech Measurements." Unpublished mearsurements, 1985.

133. Fant, G., "Analysis and Synthesis of Speech Processing." In B. Malmberg (Ed.), Manual of Phonetics. Amsterdam: North-Holland Publishing Company, pp 173-227, 1968.

134. Fant, G., The Acoustic Theory of Speech Production. The Hague: Mouton. 1970.

12. VITA

Biographical items on the author of the dissertation, Mr. H. M. Chang

1) Born January 11, 1954.

2) Attended the Nanking University in Nanking, China, from September, 1973 to January, 1977. Received the degree of Bachelor of Science in Computer Science in January, 1977.

3) Product Testing Engineer, Nanking Communication Corporation in Nanking, China, January, 1977 to August, 1978.

4) Graduate Research Fellow, the Institute of Computing Technology of the Chinese Academy of Science, Peking, China, September, 1978 to September, 1980.

5) Attended the Graduate School of the Chinese Academy of Science in Peking, China, from September, 1978 to September, 1980.

6) Teaching Assistant, the Department of Mathematics at Memphis State Univeristy, January, 1981 to December, 1981.

7) Attended Memphis State University from January, 1981 to December, 1981. Received the degree of Master of Science in Mathematics in December, 1981.

8) Research Assistant, the Center for Computing System Design at Washington University in St. Louis, May, 1982 to May, 1983.

9) Research Assistant, the Central Institute for the Deaf in St. Louis, January, 1983 to September, 1986.

10) Attended Washington Univeristy in St. Louis from January, 1982 to December, 1986.

11) Member of Technical Staff, the Advanced Technology Development at NYNEX Corporation, September, 1986 to the present date.

12) Membership in Professional Societies:
Association of Computing Machinery, American Society of Acoustics and American Association for Artificial Intelligence.

May, 1987

Short Title: SWIS: A Word Recognition System Chang, D.Sc. 1987

JDM

WASHINGTON UNIVERSITY

SEVER INSTITUTE OF TECHNOLOGY

Classifying Speech into Silence, Glottal Source, Burst Friction, or Mixed Categories

by

Steven J. Sadoff

Prepared under the direction of Dr. J. R. Cox and Dr. J. D. Miller

---

A dissertation presented to the Sever Institute of

Washington University in partial fulfillment

of the requirements for the degree of

Doctor of Science

May, 1990

Saint Louis, Missouri

AFOSR Grant G-AFOSR-86-0335
Final Technical Report
Appendix

B

WASHINGTON UNIVERSITY

SEVER INSTITUTE OF TECHNOLOGY

ABSTRACT

Classifying Speech into Silence, Glottal Source, Burst Friction, or Mixed Categories

by Steven J. Sadoff

ADVISORS: Dr. J. R. Cox and Dr. J. D. Miller

May, 1990

Saint Louis, Missouri

The primary goal of this research is to automatically segment speech into four acoustic categories based on the location of the sound sources in the human vocal tract: silence (S), glottal source (G), burst friction (B), mixed (M). A multidisciplinary approach has been taken in an attempt to solve this interesting and important problem in speech processing. In addressing this problem, knowledge and techniques from many fields including, but not limited to digital signal processing, artificial intelligence, pattern recognition, neural networks, psychophysics, speech perception, speech production, and the acoustics of speech have been applied. A connectionist system has been implemented and trained on 34 seconds of continuous speech from a female speaker. When tested on an additional 87 seconds of speech from the same speaker, the classification was 90.0 percent correct and when tested on 106 seconds of speech from a male speaker the accuracy was 88.8 percent.

TABLE OF CONTENTS

1 Introduction 1

1.1 Description of the Four Categories 2

1.2 Definition of the Problem 6

1.3 General Approach Towards Solving the Problem 7

2 Motivation and Review of Related Work 8

2.1 Motivation 8

2.2 Review of Related Work 10

2.3 Endpoint Detection 12

2.4 Silence Detection and Voice Activation 12

2.5 Periodic/Aperiodic Classification 14

2.6 Voiced/Unvoiced/Silence Classification 14

2.7 Voiced/Unvoiced/Mixed Classification 16

2.8 Other Classification Schemes 17

3 Methods 19

3.1 Algorithm Screening Database 19

3.2 Recordings 22

4 Analysis 24

4.1 Preprocessing 26

4.1.1 Digital Notch Filter 26

4.1.2 Preemphasis 27

4.1.3 Windowing 29

4.2 Information Bearing Parameters 34

4.2.1 Energy Measurements 35

4.2.2 Zero Crossing Rate 36

4.2.3 Reversal Rate 38

4.2.4 Autocorrelation 39

4.2.5 Cepstrum 39

4.2.6 Linear Prediction Error 43

4.2.7 Auditory-Based Frequency Information 45

4.3 Class Assignment 49

4.3.1 Advantages of a Connectionist Framework 50

4.3.2 The Neural Network's Perspective of the Problem 52

4.3.3 Neural Network Architecture 54

4.3.4 Training set presentation 60

4.4 Postprocessing 64

5 Implementation 67

5.1 Sampling and Displaying a Speech Waveform 67

5.2 Graphics 69

5.3 Preprocessing and Feature Extraction 70

5.4 Neural Network Simulator 72

5.5 Postprocessing 74

6 Measurements, Results, and Discussion 75

6.1 Measurements 75

6.2 Performance Criteria 80

6.3 Results 84

6.4 Analysis of the Results 89

6.5 Examination of the Neural Network 92

6.6 Effects of Postprocessing 93

6.7 Unresolved Issues and Future Directions 95

7 Summary 98

8 Acknowledgments 100

9 Appendices 101

10 Bibliography 103

11 Vita 114

LIST OF TABLES

4.1 Neural network output to real world label mapping. 59

6.1 Class Composition for the Training Sets 76

6.2 Number of segments in the training sets 76

6.3 Durations of each segment for speaker JH 77

6.4 Durations of each segment for speaker JW 77

6.5 Means and standard deviations for the four classes using the JH training data 78

6.6 Means and standard deviations for the four classes using the JW training set 80

6.7 This simplest error weighting scheme (SIMPLE) 81

6.8 Error weighting scheme based on Hamming distance (HAMMING) . . 82

6.9 Error weighting scheme for silence detection 82

6.10 Pairwise agreement over 6.900 seconds of signal 83

6.11 Confusions between speech scientists for 39.450 seconds of signal . . . 84

6.12 Performance for network NETJH 85

6.13 Performance for network NETJW 86

6.14 Confusions for NETJH RAW with JH training data 86

6.15 Confusions for NETJH POST with JH training data 86

6.16 Confusions for NETJH RAW with JH testing data 87

6.17 Confusions for NETJH POST with JH testing data 87

6.18 Confusions for NETJH RAW with JW data 87

6.19 Confusions for NETJH POST with JW data 87

6.20 Confusions for NETJW RAW with JW training data 88

6.21 Confusions for NETJW POST with JW training data 88

6.22 Confusions for NETJW RAW with JW test data 88

6.23 Confusions for NETJW POST with JW test data 88

6.24 Confusions for NETJW RAW with JH data 89

6.25 Confusions for NETJW POST with JH data 89

6.26 Effects of postprocessing on the number of segments for a single utterance spoken by subject JH 96

LIST OF FIGURES

1-1 Diagram of Sound Source Locations in the Human Vocal Tract. . . . 3

1-2 Simple Source/Filter Model for Speech Production. 5

4-1 Block Diagram of Sound Source Classification 25

4-2 Frequency Response for the 60 Hz Digital Notch Filter 28

4-3 Frequency Response for Preemphasis with $\alpha = 0.98$. 30

4-4 Spectral and Cepstral Analysis for a Glottal Source Sound. 42

4-5 The Classification Problem From the Neural Network's Perspective. . 53

4-6 Topology of the Neural Network. 58

4-7 Training Set Presentation Function. 63

6-1 Two Spiral Problem from Lang and Witbrock 1988. 79

6-2 Effects of Postprocessing on the Classification. 94

Classifying Speech into Silence, Glottal Source, Burst Friction, or Mixed Categories

1. INTRODUCTION

A fundamental problem in speech analysis is the segmentation of a speech signal into a succession of sound units with distinctly defined boundaries. To date, there is no consensus as to what the optimal unit is for speech analysis, although many different types of units have been proposed. Most of these units can be differentiated based upon the level of speech analysis that is utilized. Frequently examined levels for speech segmentation include acoustic, articulatory, phonemic, and perceptual levels. Although this hierarchy can be explicitly defined, it is not strictly layered, since there is much interaction between the different levels.

This research is an attempt to automatically segment speech into four acoustic categories based on the location of the sound sources in the human vocal tract: silence (S), glottal source (G), burst friction (B), mixed (M). Although many researchers have shown success with work on similar and related problems, none have been able to successfully address this specific problem.

A multidisciplinary approach has been taken in an attempt to solve this interesting and important problem in speech processing. In addressing this problem, knowledge and techniques from many fields including, but not limited to digital signal processing, artificial intelligence, pattern recognition, neural networks, psychophysics, speech perception, speech production, and the acoustics of speech have been applied.

A brief review of the basics of speech production will illustrate how the four categories of speech sounds differ.

1.1 DESCRIPTION OF THE FOUR CATEGORIES

Speech sounds are formed by regulating and manipulating the air stream that passes from the lungs to the external atmosphere. Although there are some languages where some of the sounds are produced during inhalation, predominantly sounds are produced during exhalation. In normal American English all speech sounds are produced by making the exhaled stream of air audible.

There are two primary locations where this air stream is used to generate sound (see Figure 1-1). The first usually involves vocal chord action. By rapidly opening and closing the elastic vocal folds, the air stream is divided into tiny puffs of air. This semi-periodic series of puffs creates a buzz like sound whose fundamental frequency is controlled by the vocal fold vibration rate. Since the orifice between the vocal folds is called the glottis, sounds created in this fashion are referred to as glottal-source sounds or voiced sounds (i.e., eat).

In the second way to make the air stream audible, a constriction is formed at a supra-glottal position along the vocal tract. Whenever the volume velocity exceeds a

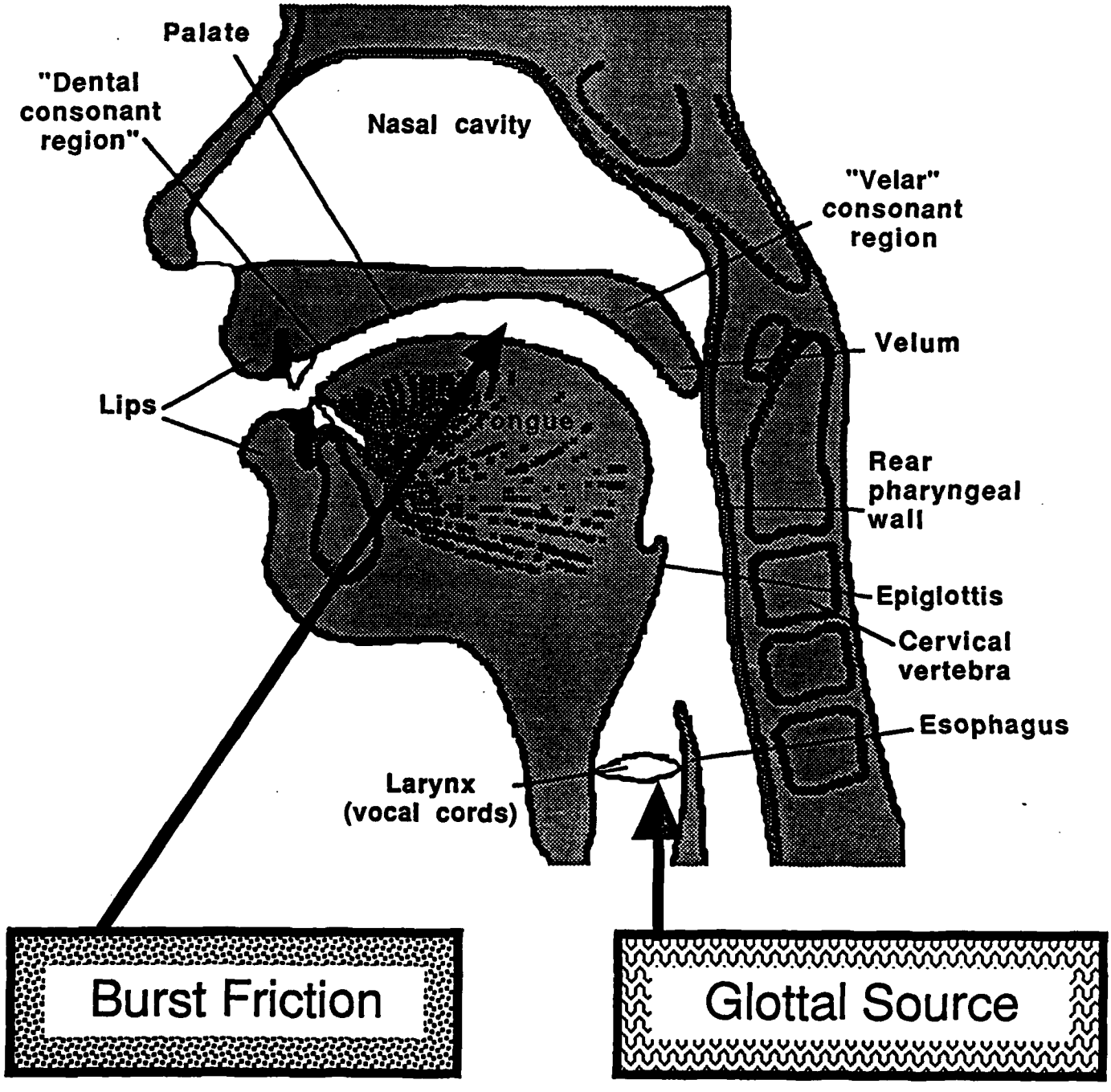


Figure 1-1: Diagram of sound source locations in the human vocal tract. After P. Lieberman and S. E. Blumstein, "Speech Physiology, Speech Perception, and Acoustic Phonetics", 1988.

critical value [1]*, the air passing this constriction becomes turbulent resulting in a hiss or burst of sound (i.e., sat). These aperiodic signals are referred to as supra-glottal or burst-friction sounds. For the frequently studied class of sounds called plosives or stops (i.e., pat) the oral cavity is completely blocked for a short period of time and usually followed by the sudden release of the built up air pressure. During the closure period, since the air flow has been interrupted, a silent segment is produced in the speech signal. If the supra-glottal occlusion is rapidly released, a small aperiodic burst or transient will occur in the signal. Finally, the two sound sources (glottal source and burst friction) can be combined simultaneously. This last class of sounds is referred to as mixed (i.e., zoo).

As described above, the four categories of sounds (S/G/B/M) are distinguished by the location of the sound sources. This particular classification scheme for identifying the location of the sound source is well summarized in Fant's classical work [2] involving speech production. To complete the story, it should be pointed out that speech production is the result of sound sources being filtered by the vocal tract. The link between these two components (source and filter) is weak and it is common to treat these as independent. A block diagram of the source model is shown in Figure 1-2. A discussion of the variables used in this figure is postponed until section 4.2.5. The filtering is controlled by varying the shape and location of the articulators (i.e., jaw, lips, tongue, pharynx, and palate). By dynamically changing the shape of the vocal tract, the resonances or formants are altered, producing different sounds. A

*The numbers in brackets in the text indicate references in the Bibliography.

more detailed description of speech production is presented by Borden and Harris [3, chapter 4] or in Flanagan's excellent book [4].

1.2 DEFINITION OF THE PROBLEM

The goal of this research is to build a system that can successfully solve the problem described in this section. First, some definitions will be presented to allow a precise description of the task to be solved. Sound can be defined as variations of pressure patterns in air that are capable of being sensed by the human auditory system. Speech is defined as air pressure variations used for the linguistic transfer of information. Natural speech is speech that is produced by a human, whereas synthetic speech is produced artificially by a machine.

The pattern classification task to be solved is formulated as a mapping problem. The domain consists of pressure variations (patterns) created by the production of natural speech. The range consists of four categories defined by the location of the sound sources in the human vocal tract and are referred to as silence (S), glottal source (G), burst friction (B), and mixed (M). Since whispered speech is produced with a constriction near or at the glottis, it is considered glottal source.

A mapping is to be performed once every millisecond (ms) and the one ms portion of the signal that the category corresponds to is referred to as a section. There is no limitation on the duration of the input signal that can be utilized in the mapping procedure and typically more than one section of the signal is used as input to the mapping procedure. The particular portion of the input signal that is used for

each classification is referred to as a frame. An important item to note, is that the classification is made solely based on the acoustic signal; no auxiliary information is used.

1.3 GENERAL APPROACH TOWARDS SOLVING THE PROBLEM

The system that is used to address the problem can be divided into four stages. In the first two stages the signal is preprocessed and 14 information bearing parameters are extracted. In the third stage, based on the information bearing parameters, a neural network is used to classify the input into one of the four classes. Finally, the stream of classifications passes through an "error correction" filter to correct any obvious mistakes. One of the more novel aspects of this approach is the use of a connectionist classifier. This work supports the findings of other researchers [5] that neural networks can be successfully applied to many problems involving speech processing.

2. MOTIVATION AND REVIEW OF RELATED WORK

2.1 MOTIVATION

There are many potential benefits and applications of an automatic algorithm for making the S/G/B/M classification. A few of the more interesting examples will now be enumerated.

One possible application of such an algorithm is the classification of speech segments into broad categories, as employed by many automatic speech recognition systems [6, 7, 8, 9, 10, 11]. For example, Reddy [6] groups segments into four nonmutually exclusive subsets: stop-like sounds, fricative-like sounds, nasal-liquid-like sounds, and vowel-like sounds. Specifically, in a multi-pass speech recognition system, if the class of a segment of speech is known, the search space can be effectively pruned reducing the complexity of the remaining passes. Groupings like these have also found application in devices for the hearing impaired. Upton [12] describes a wearable eyeglass device to assist deaf subjects in lip-reading. The eyeglasses have five miniature lights corresponding to the friction, voicing, stop, voiced-stop, and voiced-friction speech classes.

For effective high quality speech synthesis, this type of four way classification (S/G/B/M) is needed [13, 14, 1]. By implication, this classification is useful for analysis-synthesis systems. Analysis-synthesis systems are a class of systems that

estimate synthesis parameters and then use these parameters to synthesize a replica of the original speech waveform. One typical application is bandwidth reduction since only the speech parameters and not the whole speech waveform need to be transmitted [15]. Additionally, analysis-synthesis systems can be used to iteratively estimate speech parameters by analyzing the difference between the input speech signal and a synthesized version using the current estimate of the parameters [16, 17].

Another problem related to the S/G/B/M classification, is endpoint detection. In isolated word recognition systems (words are assumed to be proceeded and followed by silence), the process of locating the beginning and ending of a word is called endpoint detection. This process can be considered a subset of the S/G/B/M classification problem. Therefore, many systems, specifically isolated word recognition systems, could benefit from this project. A system that could more accurately determine endpoints would be extremely useful, since inaccurate detection of endpoints is a major source of errors for isolated word Dynamic Time Warping (DTW) systems [18].

A successful algorithm has been developed using Rate of Rise (ROR) to distinguish plosives from fricatives [19, 20, 21]. One of the major limitations, that must be solved before this algorithm can be practically applied to running speech, is that it assumes a priori that the speech segment is either a plosive or a fricative. Hence, it assumes that the signal does not contain silence, or only glottal source components (it is either a burst friction sound, or a mixed sound). Therefore, a system which could identify speech segments that fall into the proper classes, would facilitate the application

of the ROR algorithm to continuous speech. Additionally, the ROR algorithm can improve its performance (from 95.8 to 96.8 percent) if it is known whether the signal is mixed or solely burst friction.

Another application that could benefit from this project is the implementation of J. D. Miller's Auditory Perceptual Theory of phonetic recognition [22, 23]. In the first stage of his theory, the acoustic waveform is converted into sensory variables that are representative of the waveform's short-term spectral pattern. Because the characterization of the spectrum depends on the class it is in, the particular method that is used to obtain these variables is fundamentally dependent upon the four way (S/G/B/M) classification. Currently, the classification must be performed manually by a trained speech scientist. By removing the necessity for the cumbersome hand labeling, this research will facilitate the automatic implementation of Miller's Auditory Perceptual Theory.

2.2 REVIEW OF RELATED WORK

Although many researchers have investigated similar and related topics, none have tried to develop a model to distinguish these four classes of sounds based solely on the acoustic signal. As will become clear in the review that follows, there have been many attempts to solve different subsets of this problem, but it appears that these existing solutions can not be readily extended to solve the more general S/G/B/M classification. Note that removal of one or more of the four classes significantly alters the difficulty of the problem and the general utility of the solution. For example,

omitting the silence category has implications even for isolated word recognition. At first it might seem ironic, but silence plays an important role in the perception of speech. Not only is silence an important cue [24], but it has been shown that at the phonemic level, the duration of a silent interval can determine the identity of a sound [25, 26].

Most of the existing systems are based on two approaches. The first one assumes the variables of interest have particular distributions and tries to estimate the parameters that characterize the distributions. The segment is then assigned to a particular class based on the minimum computed distance using some predefined distance metric. The second approach is to develop an expert system to solve this pattern recognition task. Although this approach is generally more ad hoc, the results obtained are no worse than the more formal statistical approach.

Since almost all of the related work has been application driven, most individual systems tailor the definitions of the classes to suit the particular task at hand. Generally, this involves collapsing multiple classes or evaluating class membership less frequently (i.e., every 20 ms instead of every 1 ms). These types of simplifications dramatically reduce the complexity of the problem, allowing the development of systems with good performance.

Some closely related topics and examples of attempts to address them will now be enumerated.

2.3 ENDPOINT DETECTION

The first topic is endpoint detection, where the goal is to accurately identify the starting and ending point of a speech signal. Typically, most systems set a loudness threshold such that whenever this threshold is traversed a simple set of rules are used to verify if an endpoint has been detected [18, 27]. The basic difficulties are transients associated with either the speaker or the recording environment. Usually to overcome these problems, a set of simple features are extracted and rules are developed to discard the transient segments. Since these systems are not concerned with locating silences in the middle of speech segments, these inter-syllabic silences are considered part of the speech signal. Performance of these systems is difficult to compare not only because of the different vocabularies and different recording conditions used to test systems, but because many different schemes are used to judge accuracy. For telephone quality speech, Savoji's reported results [18] of 68% accurate, 25% acceptable and 7% fair seem fairly representative of a state of the art system.

2.4 SILENCE DETECTION AND VOICE ACTIVATION

Silence detection and voice activation are two related procedures that are typically used when a high degree of accuracy is not needed. Usually because of the nature of the applications involved, these systems must work in real time. One common application is a voice-activated microphone. Another specific need for a silence detector, arose in the early 1960s because of the high cost of long submarine cables.

Since in normal telephone conversations, each party speaks for only roughly 40% of the time, an effective silence detector can more than double the message capacity for a transmission link. To address this problem a system called Time Assignment Speech Interpolation (TASI) was developed by Bell Laboratories [28, 29]. The TASI system allowed the telephone company to more than double its transatlantic cable capacity, while having minimal or no noticeable effect on the quality of phone calls. The basic principle is that a switch is activated at a low energy threshold and remains on for several hundred milliseconds after the signal falls below the threshold. More recent efforts to develop better systems rely on more sophisticated schemes to detect the onsets of "talkspurts". For example, using a probabilistic analysis of the ratio of instantaneous to root mean square (rms) energy for both speech and noise, it was determined [30] that the onset of a talkspurt can be detected if the instantaneous voltage exceeds some threshold for a minimum duration or if the desired zero crossing rate is maintained for a minimum duration. In another approach [31], it is assumed that the background noise is stationary, allowing the onset of a talkspurt to be detected if the slope of the energy envelope is changing or if the ratio of mid to low frequency energy is changing. In these types of systems many non-speech sounds, including lip smacks, saliva pops, and clicks, are classified as speech. Also, small pauses (about 150 ms) in the speech waveform are usually classified as speech. Since the definition of the speech signal used by these systems differs from most conventional definitions, including the one used for endpoint detection schemes, the performance can not be directly compared with models developed for other applications.

2.5 PERIODIC/APERIODIC CLASSIFICATION

Classification into periodic (P) and aperiodic (A) classes is a relatively straight forward procedure that is utilized by many systems. In one of the earliest attempts to solve the P/A decision problem, Gold [32] tried to distinguish between buzz sounds (P) and hiss sounds (A) by combining the output of 6 pitch extractors. Unfortunately, the long window length (30 to 50 ms) inherent in most P/A classifiers implies that they can not track rapid changes in speech from one class to another, resulting in poor performance at class boundaries. Although all pitch detection systems need to make the voiced/unvoiced decision, most of the schemes replace it with the simpler P/A decision. Frequently, this P/A decision is based on the amplitude of the largest peak in the cepstrum [33]. Unfortunately, voiced speech is only approximately periodic [34]. Specifically, during voicing both rapid articulatory movement and the idiosyncrasies exhibited in the vocal vibrations can produce signals which are not periodic. There are two main types of variations in the glottal pulse: jitter (duration) and shimmer (amplitude). Therefore, although the P/A classification and the voiced/unvoiced classification are similar in nature, there are subtle distinctions that distinguish the two problems.

2.6 VOICED/UNVOICED/SILENCE CLASSIFICATION

One of the more frequently addressed subsets of the S/G/B/M classification problem is to classify speech into one of three categories: voiced (V), unvoiced (U), and silence (S). The mixed class (M) is considered unvoiced therefore reducing the complexity

of the more general four way classification scheme. Except for whispered speech, the G and V classes are identical and the B and U classes are identical. Although the number of studies involving whispered speech are few, the recent work by Tartter [35] shows that "Aside from periodic pulsing, spectrographic analysis of whispered speech reveals preservation of most of the properties considered important in normal speech perception". During whispered speech, the vocal cords are not vibrating. Since these sounds are produced by making the air flow turbulent (U) at or near a constriction at the glottis (G), it was decided upon to label all whispered segments as glottal source sounds.

Atal and Rabiner [36] use a pattern recognition approach to solve the V/U/S classification problem. They divide the signal into 10 ms frames, and for each frame a classification decision is made based upon the following five measurements: energy, zero-crossings, autocorrelation coefficient at unit sample delay, the first LPC coefficient (equal to the negative cepstrum at unit sample delay), and energy of the prediction error. Assuming these parameters form a multidimensional Gaussian distribution, they use the minimum non-Euclidean distance to classify each segment. This analysis is followed by a correction algorithm to "make the results appropriate for experiments in continuous digit recognition".

Rabiner and Sambur [37] use a small set of rules involving a non-linear distance measure based on a log energy computation and the first 3 poles of an 8 pole LPC analysis to classify speech into V/U/S categories. Although the system only obtains one label for every 15 ms of signal (1.5 seconds of speech would have 100 labels), their

error rate of roughly 5 percent for multiple speakers with telephone line speech is still quite impressive.

Another innovative approach for segmenting speech into one of three categories (V/U/S) was investigated by Un and Lee [38]. Their system uses bit alternations per 5 ms frame from a 1 bit linear delta modulation of the signal and zero crossings of a band pass filtered (1098 Hz to 2647 Hz) output. The basic concept is that voiced speech will only have a few alternations, unvoiced speech will have a moderate amount, and silence will have many alternations. Using a three state finite state machine (one state per class) they achieve error rates of about 3% in a noise-free environment and 6% when using recordings from a noisy environment (a computer room). Their paper is ambiguous about the specific testing conditions and there is no breakdown of what types of errors are made. They state that "we have done computer simulations with various male and female speech (English and Korean) ... The total length of the test sentences was about 25 s". Since they did not specify the number of speakers, the specific vocabulary used, or the proportions of Korean versus English, it is not clear how generally applicable their results might be.

2.7 VOICED/UNVOICED/MIXED CLASSIFICATION

In the V/U/M classification problem, the signal definitions have been altered to substantially simplify the more general problem by removing the silence category. Using a binary decision tree structure with Bayesian classifiers at each node, Siegel and Bessey [39] identify 14 features out of a possible 19 that are necessary to make the

V/U/M classification. Their system achieves 95% accuracy on speaker dependent tests and 94% accuracy on speaker independent tests. The major enhancement over the simpler V/U classification is the use of features that are correlated with the presence of mixed excitation. Because of the size and complexity of their feature set one of the clearly stated pitfalls with their work is that "For practical use, this is not computationally feasible ...". In the last eight years, technology has advanced tremendously. Enough, so that although it might be expensive, a real time system based on their approach could be built with dedicated VLSI hardware.

2.8 OTHER CLASSIFICATION SCHEMES

To solve the V/U problem, Knorr [40] presents an easily implemented method based on spectral energy distributions. Although somewhat different criteria were used in defining the signal, this system seems to have an acceptable error rate of roughly 5 percent. Another approach is presented by Kasuya and Wakita [41] for segmenting speech into vowel and nonvowel-like intervals. Their automatic system used four measurements from the signal and averaged 93 percent correct when tested on ten sentences.

In Chang's dissertation [42], a procedure, called Algorithm G, is presented to distinguish three classes of sounds, silence (S), glottal source (G), and unknown (*). Motivated by the desire to extract vocalic segments, this algorithm uses log energy, zero-crossings, the amplitude (dB) of the first peak in the spectrum, and the amplitude (dB) and frequency location (Hz) of the maximum energy peak. This algorithm yields

promising results although it has only been tested on an extremely small (the digits) vocabulary.

3. METHODS

3.1 ALGORITHM SCREENING DATABASE

An algorithm screening database has been collected and it has been used for both designing and testing all of the algorithms that were developed. Since all databases must be finite, the sampling method used for obtaining the items was carefully thought out and directly reflects the primary goals of the project. For example, if speaker independence has a higher priority than vocabulary size, then the database should have a large number of speakers at the expense of having a smaller vocabulary. The major parameters in collecting a speech database are the number of speakers, the recording environment, language and dialectical differences, whether nonsense or real words are used, whether isolated words or continuous discourse is used, the phonetic composition, and the prosodic composition.

Typically, studies involving similar problems use databases with four speakers (two male and two female) each reciting 10 to 20 seconds of speech. An explanation and justification of the different approach used for this research will now be presented.

It has been my observation that, one of the most important factors regarding variability of the speech signal is the phonetic context. Additionally, I would like to demonstrate that, at least for the speakers used for this study, a system can be designed that accurately makes the S/G/B/M classification for natural speech, albiet

read from prepared written text. Hence, I have chosen to record long passages from relatively few speakers.

The database consists of continuous speech recorded in an anechoic chamber from one native midwestern female (JW) and one native midwestern male (JH), with no known history of either speech or hearing disorders. The material that was recorded contains all of the naturally occurring American English phonemes. It has been shown that prosodic cues are not only critical to human perception, but that they can be successfully utilized for speech recognition [43]. Unfortunately, because of the massive amounts of hand labeling that would be required, it is not feasible in this study to examine each of these cues separately across all phonemic contexts. A reasonable alternative is to record a full passage, as opposed to isolated words or even complete sentences, so that a broad range of many of the important non-phonemic variables are included such as rate, rhythm, pitch variability, intensity, phrasing, intonation, and stress.

Therefore, the long phonetically balanced passage, referred to as the "rainbow passage" [44, page 127], was the text that was recorded. The full text of this passage is presented in Appendix A. The recording of the male speaker is 106 seconds in duration and the recording of the female speaker is 121 seconds in duration. The entire database consumes 8.9 MB of disk space.

A manual classification has been made for each millisecond (a frame rate of 1 frame per ms) by informal listening and by detailed computer assisted examination of the signal. A discussion of what constitutes a frame is presented in section 4.1.3. I have

classified the entire database. Since the database is 227 seconds long, this consisted of making 227,000 four way classifications. On average, it took me one hour to manually classify 3 seconds of speech or about 75 hours to classify both recordings. Also, two other speech scientists classified limited portions of the database. Specifically, one scientist (JDM) classified 6.9 seconds of speech from the male speaker and another scientist (MSF) classified 39.45 seconds of speech from the male speaker.

Originally, each millisecond was to be identified as either hard or easy reflecting the difficulty of choosing a label for the particular frame. This identification was not only to be used to automatically select better training sets, but it was also to be used to facilitate a better analysis of the mistakes that are made by the classifier. Unfortunately, this procedure was very tedious and added little additional information. Almost without failure, the hard speech segments to label were the ones near boundaries between classes and all of the other segments were easy to label. It was decided that having other scientists classify portions of the database was a better and more objective method of satisfying the purpose of the easy/hard labeling.

The passage from each speaker has been divided into two data sets (training and testing). The training set consists of the first paragraph of the passage and the testing set consists of the remainder of the passage. The data from the testing set was not used in the development of the classifier. It was solely used for cross validation.

3.2 RECORDINGS

All of the recordings were made in an anechoic chamber using a low-noise microphone/ preamplifier combination (Bruel & Kjaer 4179/2660). The microphone was placed at a height equal to and 1/2 meter in front of the subject's mouth (zero degrees angle of incidence). In this setup, turbulent noise generated by the breath flow over the microphone was not a factor. Conversational speech levels were used, i.e., 60 to 65 dBA to the microphone. The microphone output was channeled directly into a Sony PCM-501ES digital audio recorder (16 bit mode) with a JVC 720 VCR serving as the storage medium.

Prior to the recording, speakers familiarized themselves with the passage to be recorded. Speakers began reading the passage while an appropriate recording level was selected. When that recording level had been set, a calibration tone was recorded for that particular subject. The calibration tone consists of a 1 kHz sine wave generated by a Hewlett-Packard 3325A synthesizer and maintained at a constant output level of 69.5 mV (equivalent to 70 dB SPL at the microphone).

The recordings were played back from the Sony recorder in analog form and properly anti-aliased before being redigitized using a Micro Technology Unlimited DigiSound-16 analog-to-digital and digital-to-analog converter and a custom hardware interface developed in-house. Digitization was performed at 20 kHz with 16-bit precision. We know, by Nyquist's theorem, that this sampling rate limits us to frequency information below 10 kHz; this provides the system with more than adequate bandwidth for speech processing. The precision of 16 bits is also more than sufficient

for high quality speech analysis. The files are stored on a VMS VAXstation 3200.

4. ANALYSIS

One of the primary goals of this project was to develop a *working* system to perform the S/G/B/M classification. Although real time implementation was not a required goal of this work, only models which have the potential to be implemented in real time were considered. This was motivated only in part by the desire to build a working system, but mainly because of the desire to build a model which is perceptually feasible. Guided by these concerns, the parameters of the model to be developed must be relatively easy to extract from the speech signal. Using a conservative approach, whenever tradeoffs needed to be made, the system was over-engineered to handle extra accuracy at the expense of increased complexity.

Since it is highly desirable to develop a system that is speaker independent, emphasis was placed upon utilizing features that do not change on a per speaker basis. No explicit effort has been given to developing a system that is insensitive to environmental variables, though, I would like to speculate that by preprocessing the signal with the proper noise reduction techniques, the work discussed here could be adapted to yield an effective system even under noisy recording conditions.

The analysis can be logically broken down into four sequential steps: preprocessing, extraction of information bearing parameters, class assignment, and postprocessing (see Figure 4-1).

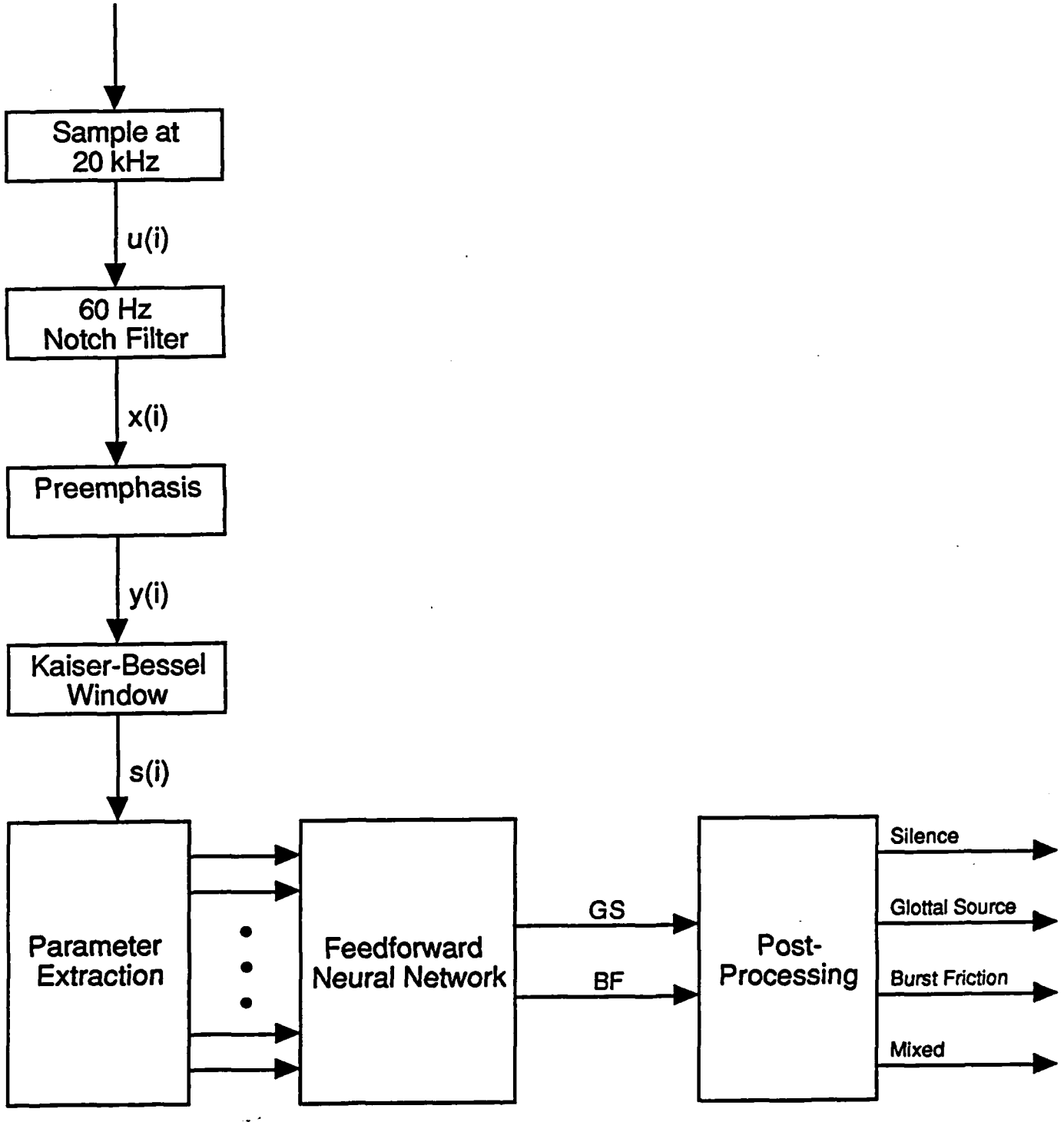


Figure 4-1: Block diagram of sound source classification.

4.1 PREPROCESSING

Preprocessing is generally some combination of simple filtering operations used both to compensate for recording artifacts such as low frequency noise or DC offset and to achieve some desired gross spectral shape.

All of the signals recorded and digitized (as specified in section 3.2) for this study have been examined for integrity and possible noise introduced in the recording and digitization process. Although, these signals do have a 60 Hz noise component (probably AC power supply noise), there seems to be minimal or no noise at harmonics of the 60 Hz noise (120 Hz and 180 Hz). Additionally, the noise around 60 Hz is constant alleviating the need for a dynamic filtering system. Other than this 60 Hz artifact, the signals seem to have no other recording artifacts.

4.1.1 Digital Notch Filter

The first step in the preprocessing stage is to filter the waveform to remove any 60 Hz components. This is done with a simple digital filter with two poles ($b = r\angle\omega_n$ and its' complex conjugate) very close to two zeros that are on the unit circle ($a = 1/\omega_n$ and its' complex conjugate). In the z domain, the transfer function is

$$H(z) = \frac{(z-a)(z-a^*)}{(z-b)(z-b^*)} \tag{4.1}$$

The discrete time function is

$$x(i) = u(i) - 2\cos\omega_n u(i-1) + u(i-2) + 2r\cos\omega_n x(i-1) - r^2 x(i-2) \tag{4.2}$$

where $u(i)$ and $x(i)$ are the ith input and output samples respectively, and the radian frequency for the zeros and poles, ω_n, is $(n2\pi)/sf$. The notch frequency, n, is set to 60 Hz, the sampling frequency, sf, is 20000 Hz, and the radius for the poles, r, is set to 0.99 (see Figure 4-2).

4.1.2 Preemphasis

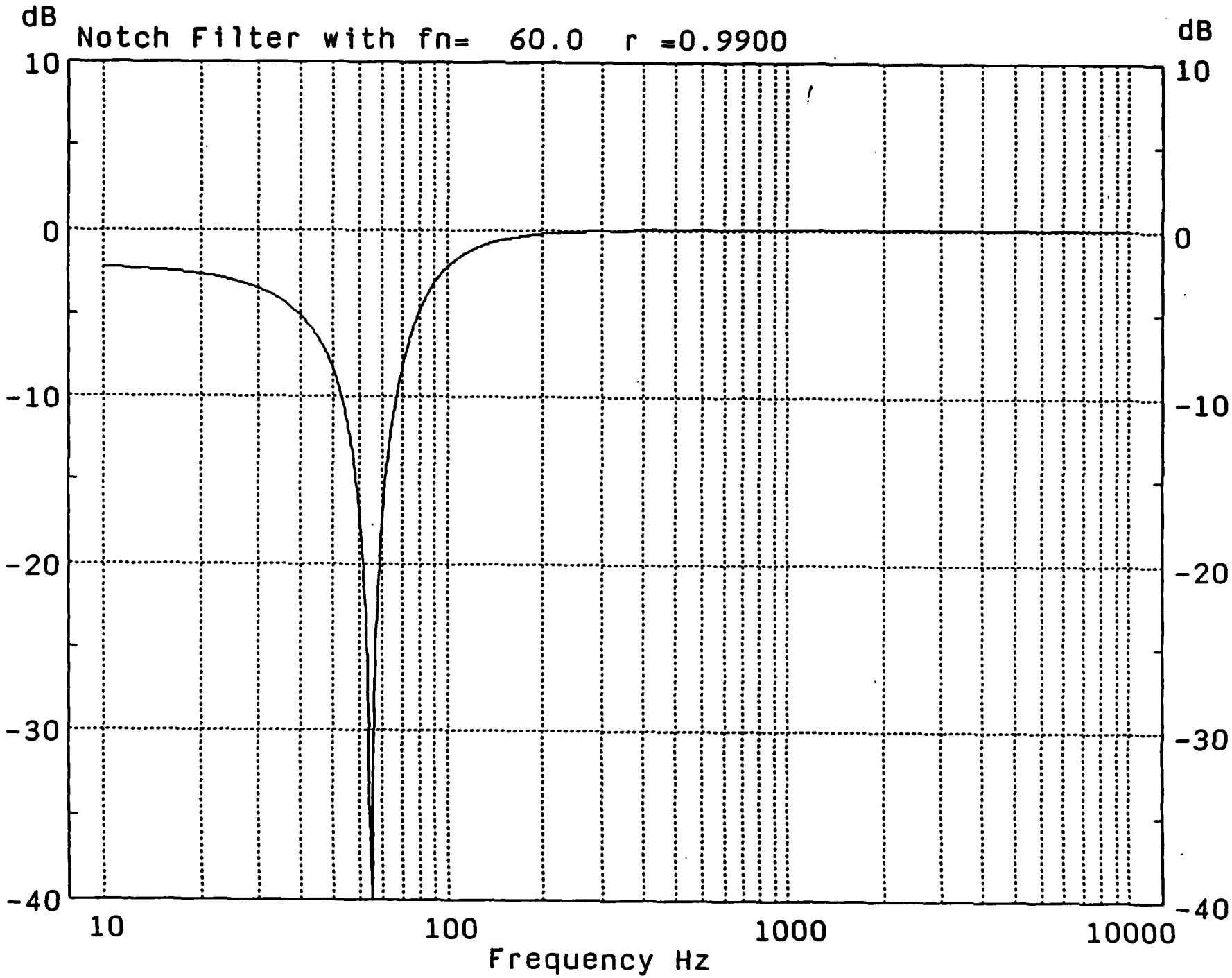
Next, the signal is preemphasized using the following first order difference equation

$$y(i) = x(i) - \alpha x(i-1) \tag{4.3}$$

In the frequency domain, this is equivalent to the following one zero filter

$$H(z) = 1 - \alpha z^{-1} \tag{4.4}$$

Acting as a differentiator, this function emphasizes the high frequency components. When trying to estimate vocal tract functions using LPC analysis, this technique is frequently used to eliminate the effects of the glottal waveform ($\approx$ -12 dB/octave) and the lip radiation characteristics ($\approx$ +6 dB/octave) [45]. Also, by flattening the spectrum, preemphasis reduces the signal's dynamic range, allowing effective fixed point LPC implementations. Using a filter-bank system for isolated word recognition, it was shown [46] that although a small improvement in average accuracy can be obtained, it was not statistically significant at the 0.9 confidence level. For vowel recognition, Paliwal showed that preemphasis was not useful, but "if the vocabulary requires more discrimination between consonants, preemphasis should be used" [47]. Since the problems of most S/G/B/M classification schemes center around the con-


dB
Notch Filter with fn= 60.0 r =0.9900
dB
10
0
-10
-20
-30
-40
10
100
1000
10000
Frequency Hz

sonants, it would seem that preemphasis should be a helpful preprocessing technique. Additionally, the speech group at CID has found it useful enough that it has become a standard procedure for all of their studies. Although the value of α does not seem to be critical, the speech group has empirically found it most useful to set $\alpha = 0.98$ which roughly corresponds to 6dB per octave. All things considered, I will use preemphasis with $\alpha = 0.98$. The frequency response of this operation is shown in Figure 4-3.

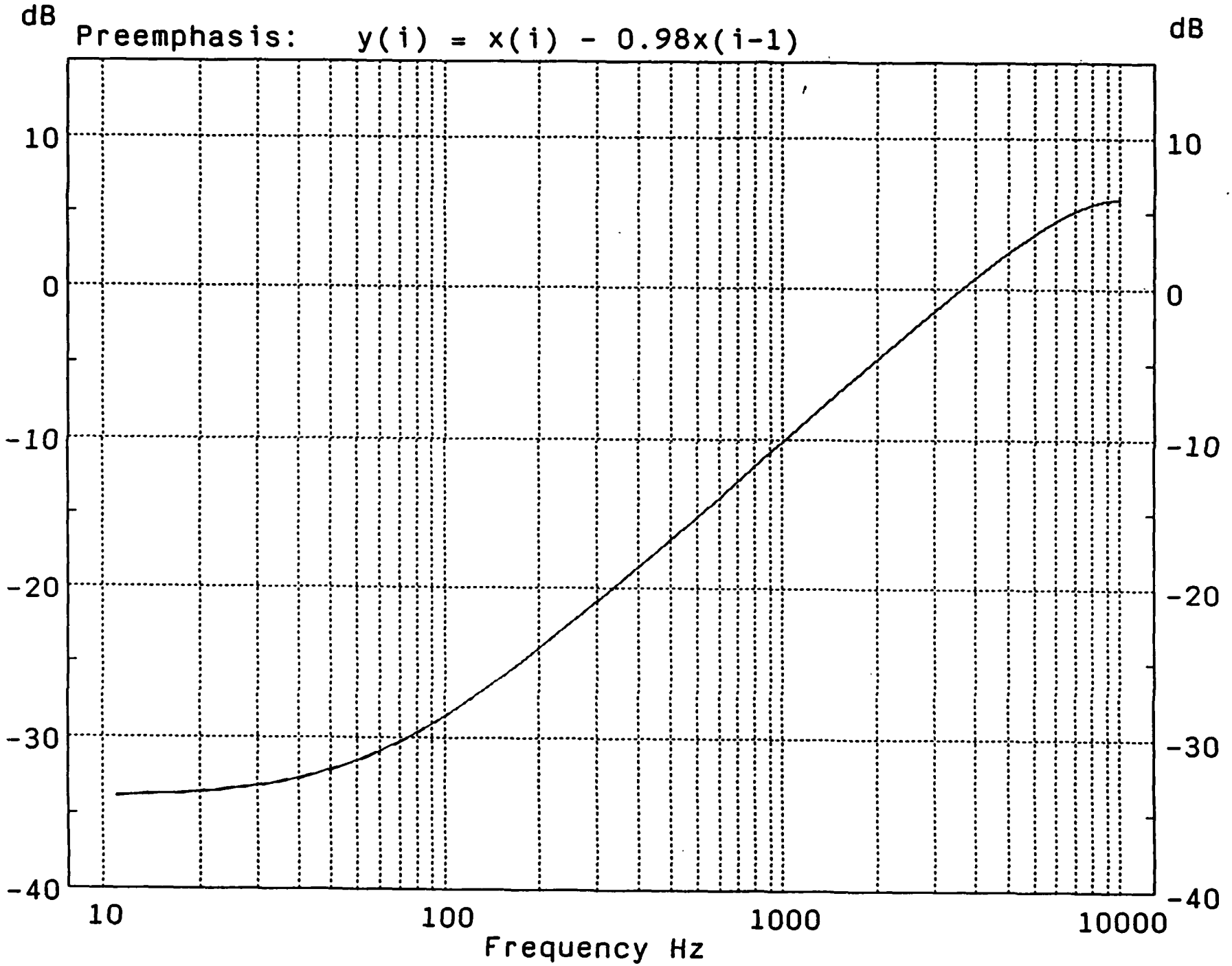
4.1.3 Windowing

Since the development of the sound spectrograph in 1946 [48, 49], time-frequency analysis of speech sounds has become commonplace. Taking advantage of the quasi-stationary nature of speech, almost all of these approaches block or window the the incoming signal into frames. This is reasonable based on the observation that the temporal and spectral properties either remain fixed over short periods of time (5 to 40 ms) or if they do change, they do so relatively slowly. There are some signals, such as a linear chirp ($e^{j\frac{1}{2}mt^2}$, the frequency is linearly increasing with time), where there is no adequate window duration. Much study has gone into choosing the appropriate window length for speech signals; the specific details and methods for blocking or windowing the signal will be discussed later after the basic problem is summarized.

It has been shown [50] by the uncertainty principle that

$$\sigma_T \sigma_\Omega \geq \frac{1}{2} \tag{4.5}$$

where σ_T and σ_Ω are variance measures indicating the degree of locality in time and


dB
Preemphasis: y(i) = x(i) - 0.98x(i-1)
dB
10
0
-10
-20
-30
-40
10
100
1000
10000
Frequency Hz

frequency respectively. This indicates that there is a fundamental tradeoff in choosing the window length for analyzing any time-varying signal (i.e., speech). Briefly stated, the limitation is that smaller window durations lead to poorer frequency responses, while larger window durations yield poorer time resolution.

To handle the concept of windowed frames, some notations need to be introduced. In general, when referring to signals, subscripts will refer to frame numbers and indices will refer to sample number. Therefore, the ith sample of the nth frame is defined as

$$x_n(i) = x(c(n-1)+i) \tag{4.6}$$

where c is the number of samples per frame shift.

Many different window shapes have been proposed for speech analysis. The basic idea behind all windowing schemes is to multiply the speech signal $x(i)$ by a finite-duration window $w(i)$ to obtain a signal that has been weighted by the shape of the window. Therefore, the final step in the preprocessing stage is the application of a window

$$s(i) = y(i)w(i) \tag{4.7}$$

A comprehensive review of applying windows (23 major families) in conjunction with a discrete Fourier transform can be found in the paper by Harris [51]. Of the many parameters in window design, two of the most important are to have a narrow main lobe and to have the side lobes reside far beneath the main lobe (in the frequency

domain). The rectangular window is the simplest example and is defined by

$$w(i) = \begin{cases} 1.0 & 0 \leq i < N \\ 0.0 & \text{otherwise} \end{cases} \tag{4.8}$$

The 6 dB bandwidth is 1.21 bins but the highest side lobe is only 13 dB below the main lobe. One of the most popular windows used for speech is the Hamming window defined as

$$w(i) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi i}{N-1}\right) & 0 \leq i < N \\ 0.0 & \text{otherwise} \end{cases} \tag{4.9}$$

The 6 dB bandwidth is 1.81 bins and the highest side lobe is 43 dB below the main lobe. The window that seems optimal for speech is the Kaiser-Bessel window defined [52] as

$$w(i) = \begin{cases} \frac{I_0\left[\pi\alpha\sqrt{1-\left(\frac{n}{N/2}\right)^2}\right]}{I_0[\pi\alpha]} & 0 \leq i \leq N/2 \\ 0.0 & \text{otherwise} \end{cases} \tag{4.10}$$

where I_0 is the modified zeroth order Bessel Function of the first kind defined as

$$I_0(r) = \sum_{k=0}^{\infty}\left[\frac{(r/2)^k}{k!}\right]^2 \tag{4.11}$$

For $\alpha = 3.0$, the 6 dB bandwidth is 2.39 bins and the highest side lobe is 69 dB below the main lobe. Therefore, in the present work, a Kaiser-Bessel window with $\alpha = 3.0$ is applied to the speech waveform after preemphasis has been performed.

As has been indicated earlier, the duration of the frame is an integral issue for any speech analysis scheme. In reality, the window duration does not even have to be finite, since an infinite impulse response (IIR) filter can be used for the window

function. Because speech is only pseudo stationery, one most be cautious before choosing a window length.

Intuitively, if the window length is long, the computed information in the time domain will change little over time, performing a lowpass filter operation. If the duration of the window is too long (i.e., more than several pitch periods) the computed information in the time domain will not reflect the changing properties of the speech signal. The shorter the duration of the window, the more responsive the computed information in the time domain will be to rapid changes. If the duration of the window is too short (i.e. less than a pitch period), the computed information in the time domain will not be a smooth function.

From Peterson and Barney [53], it is known that the geometric means for the fundamental frequency for vowels are 132 IIz, 223 Hz, and 264 Hz for men, women, and children respectively. Alternatively, the average respective pitch periods are 7.6 ms, 4.5 ms, and 3.8 ms. Therefore, a practical range for the window duration is usually between 10 and 40 ms (2 to 4 pitch periods). For spectral analysis, the speech group at CID has empirically found 24 ms (480 samples) frames to be most useful.

Since many of the cues for the S/G/B/M classification are far less than 24 ms in duration, I was hesitant about using such a long window for this task. Experimentally, I have found that a window length of 12 ms seems to yield good results. Therefore, for this project, unless explicitly noted, the window length has been fixed at 240 samples (12ms). The symbol that is used for the window length is N.

This window is moved along in 1 ms steps and this will be referred to as the frame

rate. This means that one frame is obtained for every millisecond of signal. Because the acoustic events that concern us are sometimes as short as 3 or 4 ms in duration (i.e., the burst of a voiceless plosive), and because of the desire to have accurate time localization of the acoustic events, the frame rate is fixed at 1 frame per ms for this entire research effort.

4.2 INFORMATION BEARING PARAMETERS

Many different features have been examined as potential parameters for S/G/B/M classification task. Parameters from both the time and frequency domains have been examined. In total, fourteen information bearing parameters were developed for the S/G/B/M classification task. It is presumed that the variations of these features bear the relevant information for identifying the sound source. These features seem to be a sufficient set of parameters to successfully address the problem at hand. While an accurate method has been developed for the S/G/B/M classification using a specific set of 14 information bearing parameters, various combinations of other features may yield similar results.

Generally, when trying to identify relevant features there is an implicit tradeoff that has to be evaluated. In many psychophysical tasks, it can be observed that the usefulness or informational content of a particular parameter is inversely proportional to how reliably the parameter can be extracted. For example, parameters such as formants, pitch, and loudness have high information content but are not easy to automatically extract; parameters that are less perceptually meaningful such as zero

crossings, fundamental frequency, and energy, are relatively easy to extract from the signal. Therefore, if U_a is the useful information content of some feature a and ρ_a is the accuracy with which it can be extracted, then

$$U_a \propto \frac{1}{\rho_a} \tag{4.12}$$

Additionally, if we have two features, a and b, and we can only use one of them, then we use the following rule

$$\rho_a U_a > \rho_b U_b \quad \text{choose feature } a \tag{4.13}$$

$$\rho_a U_a < \rho_b U_b \quad \text{choose feature } b \tag{4.14}$$

The information bearing parameters that are used will now be enumerated.

4.2.1 *Energy Measurements*

The short-time energy and amplitude are important parameters. It is well known that voiced sounds generally have high overall intensities while unvoiced sounds generally have lower overall intensities. Perceptual studies [54] have shown that for the phonemic voiced/voiceless distinction overall intensity is a dominant cue. The calculation of energy involves summing the squares of the input signal, while the calculation of amplitude simply sums the absolute values of the input signals. Although similar in nature, energy calculations are more popular and are commonly used for both broad categorization of speech sounds and endpoint detection (e.g., [55, 7, 27]).

It should be noted that intensity is a physical property of a signal, whereas loudness is the subjective perception of intensity. Intensity is usually measured on a

logarithmic scale using decibels (dB) and loudness is measured using a unit called a phon. Although intensity and loudness are related, the correlation is imperfect because when two sounds of different frequencies and identical intensities are presented to a human subject, the perception is not always one of equal loudness [56]. To match the performance of the human auditory system, detailed approximation procedures have been developed to try to automatically convert intensity into loudness [57].

For the S/G/B/M classification, the extra effort to include loudness as a parameter does not seem justified, and hence an easily extractable measure of intensity is used instead. Thus, the short term log rms energy for the nth frame, RMS_n, defined as

$$RMS_n = 10 \log \left((1/N) \sum_{i=1}^{N} [s_n(i)]^2 \right) \tag{4.15}$$

is used for the classification procedure. Before the energy calculation, the signal is first preemphasized and then multiplied by a 12 ms (240 samples) Kaiser-Bessel window.

4.2.2 Zero Crossing Rate

Zero crossing rate gives a general indication of the frequency location of the major energy concentration. For example, for vowel-like sounds, the zero crossing rate will follow the first formant, since it usually has the highest energy. The zero crossing rate is calculated from the raw input waveform, before preemphasis, since it has been shown [58] that for speech recognition the zero crossing rate of the raw signal is better than the zero crossing rate of the preemphasized version. Because of the recording conditions for this study (i.e., an anechoic chamber with high quality recordings), the

effects of background noise and DC offset are minimal allowing the simple calculation of the number of times the input signal crosses the axis (zero crossings) to be an effective parameter.

Clearly, applying any type of standard window does not alter the number of zero crossings since the sign of the signal is not changed. The zero crossing rate per second, ZC_n, for frame n is defined as

$$ZC_n = \frac{sf}{N-1} \sum_{i=2}^{N} |sign[x_n(i)] - sign[x_n(i-1)]| \tag{4.16}$$

where

$$sign[i] = \begin{cases} -0.5 & x < 0 \\ 0.5 & x \geq 0 \end{cases} \tag{4.17}$$

For example, for a 1000 Hz sinusiod $ZC_n = 2000$. Since the frame length is 240 samples, there can be from 0 to 239 zero crossings per frame, yielding the range of values for crossings per second of $0 \leq ZC_n \leq 20000$.

As a side note, there are many interesting perceptual studies involving zero crossings demonstrating its compactness as a coding scheme for speech. The concept behind removing the amplitude information from the signal and only keeping the sign of the signal is referred to as "infinite" peak clipping and dates back to at least 1947. One way of implementing this process is to simply digitize the signal using a 1 bit analog to digital converter. Two early studies [59, 60] demonstrated that "a considerable amount of information is carried by the temporal pattern of the crossings of the zero-axis". It was shown that if speech is sampled at 10000 Hz and then it is infinitely peaked clipped, trained listeners found the signal highly intelligible (over

97 percent word articulation). More recent studies on zero crossings have further investigated the perceptual details [61], different mathematical formulations [62], and using zero crossings as a method for formant tracking [63].

4.2.3 Reversal Rate

This is simply the number of times that the signal changes direction per frame. This can be computed by taking the zero crossing rate of the derivative of the input signal. The reversal rate per second, RR_n, for frame n is defined as

$$RR_n = \frac{sf}{N-2} \sum_{i=3}^{N} |sign[x_n(i) - x_n(i-1)] - sign[x_n(i-1) - x_n(i-2)]| \qquad (4.18)$$

where $sign[i]$ was defined in equation 4.17.

Just as for zero crossings, for a 1000 Hz sinusiod $RR_n = 2000$. But for more complex signals there are substantial differences between reversal rate and zero crossing rate. For example, for a low amplitude high frequency signal modulated by a high amplitude low frequency signal, the reversal rate would track the high frequency component while the zero crossing rate would track the low frequency component. Since the frame length is 240 samples, there can be from 0 to 238 reversals per frame, yielding the range of values for reversals per second of $0 \leq RR_n \leq 20000$.

4.2.4 Autocorrelation

The normalized autocorrelation coefficient at unit sample delay is another information bearing parameter used for the classification. It is defined by

$$AC_n = \frac{R_n(1)}{R_n(0)} \tag{4.19}$$

where $R_n(k)$ is the autocorrelation function defined by

$$R_n(k) = \sum_{i=1}^{N} s_n(i)s_n(i-k). \tag{4.20}$$

$R_n(k)$ is an even function, where $f(k)$ is an even function if

$$f(k) = f(-k). \tag{4.21}$$

Since the maximum value of $R_n(k)$ is at $k = 0$, then by definition $-1 \leq AC_n \leq 1$. The autocorrelation is performed after both preemphasis and windowing. A value of $N = 240$ (12 ms) is used. Even at unit sample delay, the correlation of adjacent samples, AC_n, is generally large for voiced speech and small (close to 0) for unvoiced speech. Although the autocorrelation can also be obtained by taking the spectrum of the power spectrum, this is not utilized in the current implementation.

4.2.5 Cepstrum

In 1959, Tukey suggested that the power spectrum of the logarithm of the power spectrum might be useful for calculating the difference in arrival times of seismic echos for determining the focal depth of seismic events. In 1960, at Bell Telephone Laboratories, Bogert [64] programmed Tukey's suggestion to analyze earthquakes and

explosions. Tukey's terminology was a play on words; "cepstrum" was the spectrum of the log spectrum and the frequency of the spectral ripples were called "quefrency". Unfortunately for Bogert and his colleagues, the cepstral analysis of seismic events did not turn out to be promising. It took until June 1962, before someone (Schroeder) suggested using cepstral analysis for pitch determination. An important lesson to be learned, is that although a particular technique might not be useful in one domain, it should not be casually discarded since it might prove useful for other classes of problems.

The cepstrum has become a frequently used method for measuring pitch [33], since for periodic signals it has a strong peak corresponding to the pitch period. For a particular frame, i, the cepstrum is defined as

$$C_i(n) = \frac{1}{N} \sum_{k=0}^{N-1} \log |S_i(k)| e^{j\frac{2\pi}{N}kn} \quad n = 0, 1, 2, \ldots, N-1 \tag{4.22}$$

where $S_i(k)$ is the Discrete Fourier Transform defined as

$$S_i(k) = \sum_{n=0}^{N-1} s_i(n) e^{-j\frac{2\pi}{N}kn} \quad k = 0, 1, 2, \ldots, N-1 \tag{4.23}$$

A review of how the cepstrum separates the effects of the vocal tract and the sound source will now be presented.

First let us review the simplified model of speech production shown in Figure 1-2 and define some of the variables involved. The source signal is $g(i)$, it's spectrum is $G(k)$, the impulse response of the vocal tract is $h(i)$ and its spectrum is $H(i)$. The output speech signal is $u(i)$ and it's spectrum is $U(k)$. Also, the $*$ operator denotes the convolution operation and the $\mathcal{F}$ denotes the Fourier transform.

The basic source/filter model tells us that

$$u(i) = g(i) * h(i) \tag{4.24}$$

In the frequency domain this is equivalent to

$$U(k) = G(k)H(k) \tag{4.25}$$

This means that the speech spectra is equal to the product of the source spectra and the spectra of the vocal tract. If we take the log magnitude of both sides of equation 4.25 and apply the simple algebraic fact that

$$\log(ab) = \log(a) + \log(b) \tag{4.26}$$

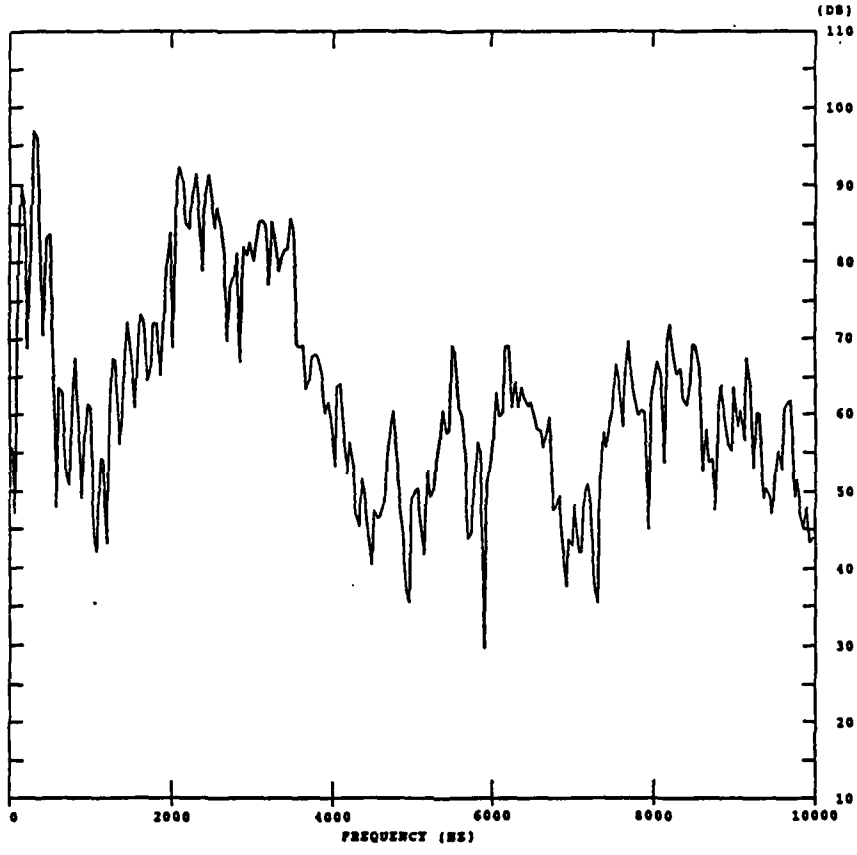
we obtain

$$\log|U(k)|^2 = \log|G(k)|^2 + \log|H(k)|^2 \tag{4.27}$$

Since addition is preserved in both the time and frequency domain, we can take the Fourier transform of both sides of this equation and obtain

$$\mathcal{F}[\log|U(k)|^2] = \mathcal{F}[\log|G(k)|^2] + \mathcal{F}[\log|H(k)|^2] \tag{4.28}$$

The main effect of the sound source is a high frequency ripple (harmonics) in the log spectrum, while the major effect of the vocal tract is low frequency components in the log spectrum. Hence, in the cepstrum (spectrum of the log spectrum) a periodic sound source with period T shows up as a fairly sharp peak at T cycles per hertz (seconds). The filtering performed by the vocal tract is manifested as broad peaks with low quefrency values (less than T). In Figure 4-4, both a spectral and a cepstral



Cepstral analysis for a voiced sound

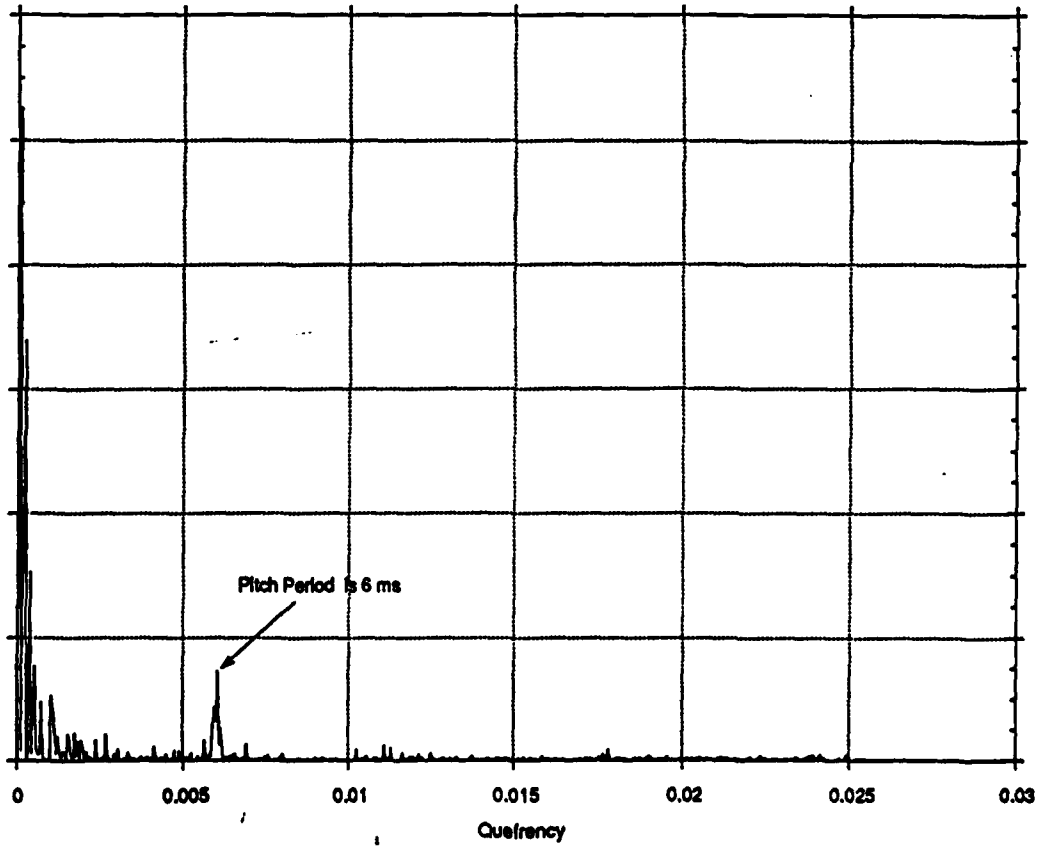


Figure 4-4

analysis are shown for a male speaker saying "IY" (the first vowel in the word these).

Because of desire to have at least one pitch period in the frame being analyzed, the standard 12 ms window is not used for the calculation of the cepstrum. Instead, a 1024 sample (51.2 ms) window is used for the cepstrum.

For voiced speech, the cepstrum frequently has a peak between 2.5 and 15 ms (corresponding to a pitch between 400 Hz and 66.7 Hz). Typically, the signal is labeled periodic if the magnitude of this peak exceeds some threshold. Because the cepstrum is generally downward sloping (smaller values at higher quefrencies), I have adapted the linear multiplicative weighting scheme suggested by Noll [33]. Specifically, the magnitude of the cepstrum is multiplied by a weight that is 2 at 2.5 ms and 5 at 15 ms.

Finally, the magnitude of the largest peak between 2.5 and 15 ms is used as an information bearing parameter for the the S/G/B/M classification.

4.2.6 Linear Prediction Error

For speech signals, one of the most popular models used to represent and estimate the basic speech parameters (e.g., fundamental frequency, vocal tract area functions, formants, low bit rate coding) is referred to as Linear Prediction Coding (LPC) [65, 45]. The popularity of this all pole (autoregressive) model is not only due to the relative ease of computation, but to the compact and precise representation of the speech spectrum. There are many equivalent formulations that have been developed independently, but the basic idea is that a speech sample can be approximated by a

linear combination of previous speech samples.

I will now present an overview of what is generally referred to as the autocorrelation method for linear prediction. We try to estimate $s(i)$ with the following equation

$$\hat{s}(i) = \sum_{k=1}^{p} a_k s(i-k) \tag{4.29}$$

where p is the number of poles (previous samples) used to estimate $s(i)$. The prediction error, $e(i)$, is given by

$$e(i) = s(i) - \hat{s}(i) = s(i) - \sum_{k=1}^{p} a_k s(i-k) \tag{4.30}$$

Summing this error over a complete frame yields

$$E_n = 1/N \sum_{i=1}^{N} e(i)^2 = 1/N \sum_{i=1}^{N} \left(s(i) - \sum_{k=1}^{p} a_k s(i-k) \right)^2 \tag{4.31}$$

The a_k are chosen to minimize this mean-squared prediction error. This is done by setting $\partial E_n / \partial a_l = 0$ for all $l = 1, 2, \ldots, p$ and then solving the p linear equations. After doing this, it is straightforward to show [45] that the a_k's can be obtained by solving the following p equations

$$R_n(l) = \sum_{k=1}^{p} a_k R_n(|l-k|) \quad 1 \leq l \leq p \tag{4.32}$$

where $R_n(l)$ is the autocorrelation defined in equation 4.20. Then it can also be shown that the error is

$$E_n = R_n(0) - \sum_{k=1}^{p} a_k R_n(k) \tag{4.33}$$

The error, also referred to as the residual, reflects how well the LPC model represents the spectrum. The information bearing parameter that is used for the classifier is

denoted $LPCR_n$, and is defined as

$$LPCR_n = RMS_n - 10\log(E_n) \tag{4.34}$$

In general, glottal source sounds will have smaller residuals (better LPC fit) than burst friction sounds. Additionally, slightly larger error signals are obtained when an LPC frame is centered on an epoch (open phase of a glottal pitch pulse) [66]. Hence, for glottal source sounds, the residual is not only smaller than for burst friction sounds, but it is also periodic.

4.2.7 Auditory-Based Frequency Information

Frequency information is an important cue for making the S/G/B/M classification. Generally, glottal-source sounds have a large amount of low frequency energy, burst-friction sounds have a large amount of high frequency energy, and mixed sounds have both high and low frequency energy. The item that seems to have the most importance is the relative amounts of high frequency versus low frequency information.

One of the more standard methods for analyzing speech is Fourier analysis. Although many new and potentially useful types of spectral analysis are constantly being developed (i.e., Frazier-Jarwerth transforms [67, 68]), I was hesitant about using these more novel methods until they have been shown to be beneficial for speech analysis. For example, although the Wigner distribution seemed promising for speech analysis, recent results [69] indicate that because of artifacts which appear between formants, it does not offer significant improvements over the spectrogram. Some people advocate the use of these techniques because they offer computational

advantages. For the S/G/B/M classification, this would be of little benefit, since the spectral analysis is not a computational bottleneck in the classification procedure.

I have chosen to use an auditory-based frequency representation. The basic idea is to use knowledge obtained from pyschoacoustics to come up with a better computational model for speech processing [70, 71, 72]. The work in this field can be traced back to 1924 when Wegel and Lane [73] reported the first systematic and quantitative study of masking.

Much effort has gone into trying to model the human auditory system, and almost all existing models use some sort of filter bank structure. A substantial effort has been devoted to determining the characteristics of this hypothesized filter bank. Issues that have been addressed are not limited to but include filter spacing, filter bandwidths, the number of filters, and the filter types. Most of these issues have been evaluated in terms of predicting human performance in particular, well-constrained psychoacoustic tasks. One of the most well examined filter bank models is the critical band.

First, let me define masking as the increase (in dB) of the threshold for the detection of one sound in the presence of another (a good review of masking is presented by Jeffress [74]). If a pure tone is presented along with a narrow band of white noise, it turns out that there is only a limited bandwidth around the tone that contributes to the amount of masking for the tone. This bandwidth is referred to as a critical band [75, 76, 77, 78]. Since the critical band surfaces in so many aspects of psychoacoustics (i.e., thresholds, pitch, loudness, masking, and musical consonance) there have been many experimental paradigms that have been developed to empirically determine the

nature of the critical band. The general finding is that the critical band is roughly constant below 500 Hz and becomes proportionally wider as the center frequency of the band increases beyond 500 Hz. The definition that I prefer most is presented in the comprehensive chapter written by Scharf [79], "As a purely empirical phenomenon, the critical band is that bandwidth at which subjective responses rather abruptly change". For the purposes of this research, a critical band describes the bandwidth of the auditory filter at a given center frequency.

It should be also be pointed out that although the critical band is ubiquitous in psychoacoustics, there is a growing amount of evidence [80, 81, 82, 83, 84] that it is a dramatic oversimplification of how the auditory system works. Since it still seems like the best overall engineering approximation for the auditory system, it is the model that I have used for spectral analysis.

Unfortunately, there have been only a few quantitative studies comparing and evaluating the relative importance and performance of filter-bank parameters for speech recognition tasks. In one of the most comprehensive studies of filter bank design for isolated word recognition [85], some important results were presented which have guided my work.

Their results were obtained using a 39 word vocabulary, 2 male and 2 female speakers, and speech that was band limited to the range of 200 Hz to 3200 Hz. The first result was that too few (3 or less) or too many filters (more than 31) for nonoverlapping filter banks degrades performance. They also showed that performance was best when the composite spectrum was relatively flat. Also, for nonuniform filter

banks, performance obtained when filters were spaced along a critical band scale was significantly better than when spaced on octave bands, 1/3 octave bands, or arbitrary spacings. For their word recognition task, the performance of LPC-based recognizer was statistically better than any of the filter banks. I have chosen not to use an LPC-based spectral model, since for wide band speech (not cutoff at 3200 Hz), this difference is probably not significant [86]. The motivating factor for this decision was the desire to use a spectral representation that closely reflects the underlying perceptual system. Additionally, since LPC based models are extremely suseptable to environmental noise, models that depend upon an LPC representation are generally restricted to excellent recording conditions.

The problem I am addressing is significantly different than isolated word recognition. The concern is that results obtained and parameters that are evaluated using isolated word recognition performance, may not carry over to the S/G/B/M classification.

In light of the many results for choosing optimal spectral coefficients I have made the following design decisions. The often used model of overlapping critical band filters seems like the best frequency representation. The log magnitude of the output of 8 overlapping critical band filters as specified by Moore and Glasberg [78] are used as information bearing parameters. These 8 parameters are referred to as $S1$, $S2$, ..., $S8$.

4.3 CLASS ASSIGNMENT

Ideally, the decision-making procedure would be separate and independent of the particular input parameters. Because this is not the case, the design of the classifier has been a highly interactive process. This interactive process could not have been as successful as it was without the rich computerized environment at my disposal. The important role of general purpose "graphics-oriented interactive pattern analysis and classification systems" is explicitly stated by Kanal [87].

One approach to solving the classification problem (referred to as a statistical model) would be to estimate the parameters for the distributions of each of the previously mentioned measurements and classify a segment based on some simple distance metric. Some of the problems with this approach are that assumptions must be made about underlying distributions and that the individual distributions show significant overlap. This approach also has the disadvantage of ignoring structural properties of the input while it focuses entirely on the statistical properties of the set of scalar information bearing parameters.

Another approach, sometimes referred to as linguistic modeling, would be to develop a set of rules (or a grammar) based on the observed values of each of the parameters. Although this approach can also yield success, much effort is usually spent hand-crafting the, sometimes ad hoc, set of rules.

4.3.1 Advantages of a Connectionist Framework

Since a connectionist framework seems ideally suited to the S/G/B/M classification, this is the alternative that was investigated. Neural networks are a class of models that perform some computation by the mere action of a large number of simple processing elements (called units) sending signals along connections to other units. Associated with each connection is a multiplicative weight that indicates the strength of the signal to be propagated. Typically, during use these weights are modified based on current results to improve performance.

Many different terms have been used for these models including, but not limited to neural networks [88], neural computers [89], connectionism [90], parallel distributed processing [91], cognizers [92], and parallel networks [93]. Because most connectionist models have been guided by simplified models of neural mechanisms, the models typically reflect a "biological flavor". One of the ultimate hopes is that the recent coevolution of both neuroscience and cognitive science will assist in explaining and understanding how the mind-brain works [94]. Although, none of the existing models present a complete account of the underlying physiology, most are dramatically closer than standard information-processing models. Of the many potential benefits of neural nets, there are a couple that are particularly useful for this study.

Conventional computers are vastly different computationally than biological systems. Biological systems have certain timing constraints that have direct implications for sequential computations. The following reasoning is presented by Feldman [95]: a back of the envelope calculation says that since neural elements operate at speeds of a

few milliseconds and since humans can solve many perceptual tasks in a few hundred milliseconds, these tasks can be solved in roughly one hundred time steps. This can be a severe limitation for many conventional artificial intelligence systems that claim to model some aspect of human perception. The connectionist system used for this work satisfies the one hundred cycle constraint.

Another useful property of connectionist systems is their ability to adapt and continue learning. If the input conditions change sufficiently, a new training phase can be performed which allows the network to adapt to the new conditions. For follow up work related to this project there are two potential uses for this property. The first reason for retraining would be to optimize performance on a per speaker basis. Secondly, retraining might be useful for different recording environments.

Since neural networks are non-parametric and make weaker assumptions about the underlying distributions, they can frequently be more powerful than standard statistical approaches. Statistical approaches make certain simplifying assumptions (i.e., linear processes or Gaussian distribution) about the underlying problem [96] that are sometimes unjustified and are used solely for the purpose of mathematical tractability. With connectionist systems, this problem is minimized.

Another property of connectionist systems that is helpful for this study is the ability to integrate multiple simultaneous constraints. Even if the inputs are from diverse sources, neural networks allow fairly uniform representations of many different inputs.

The extensional as opposed to intensional computational model is also an attrac-

tive feature of neural networks, since the ability to learn by example can be more attractive than the more conventional and time consuming approach of developing a complicated set of rules. This principle is well illustrated by a neural network system, NETtalk [93], that converts unrestricted English text to speech. Whereas most systems for text-to-speech conversion involve traditional rule programs that generally take many man years to design [97], NETtalk learns to "read aloud" by repeatedly presenting example phonetic transcriptions to the system. After the training is done, the system not only achieves good performance, but it generalizes to novel words. This remarkable system demonstrates that complicated tasks, such as text to speech conversion, are candidates for extensional learning paradigms such as neural networks.

Another reason why connectionist systems are attractive is that they provide an alternative computational language from which potential solutions can be expressed. Throughout the history of computer science it has been shown that the choice of languages has a profound effect on how problems are thought about and on how algorithms are designed. The concept that many problems are better formulated using a connectionist framework, will hopefully lead to not only more optimal solutions, but also to better insights.

4.3.2 The Neural Network's Perspective of the Problem

A review of the problem that the classifier was designed to address will now be presented. To better illustrate the task from the neural network's perspective, a dataflow description of the problem is presented in figure 4-5. A portion (26 ms)

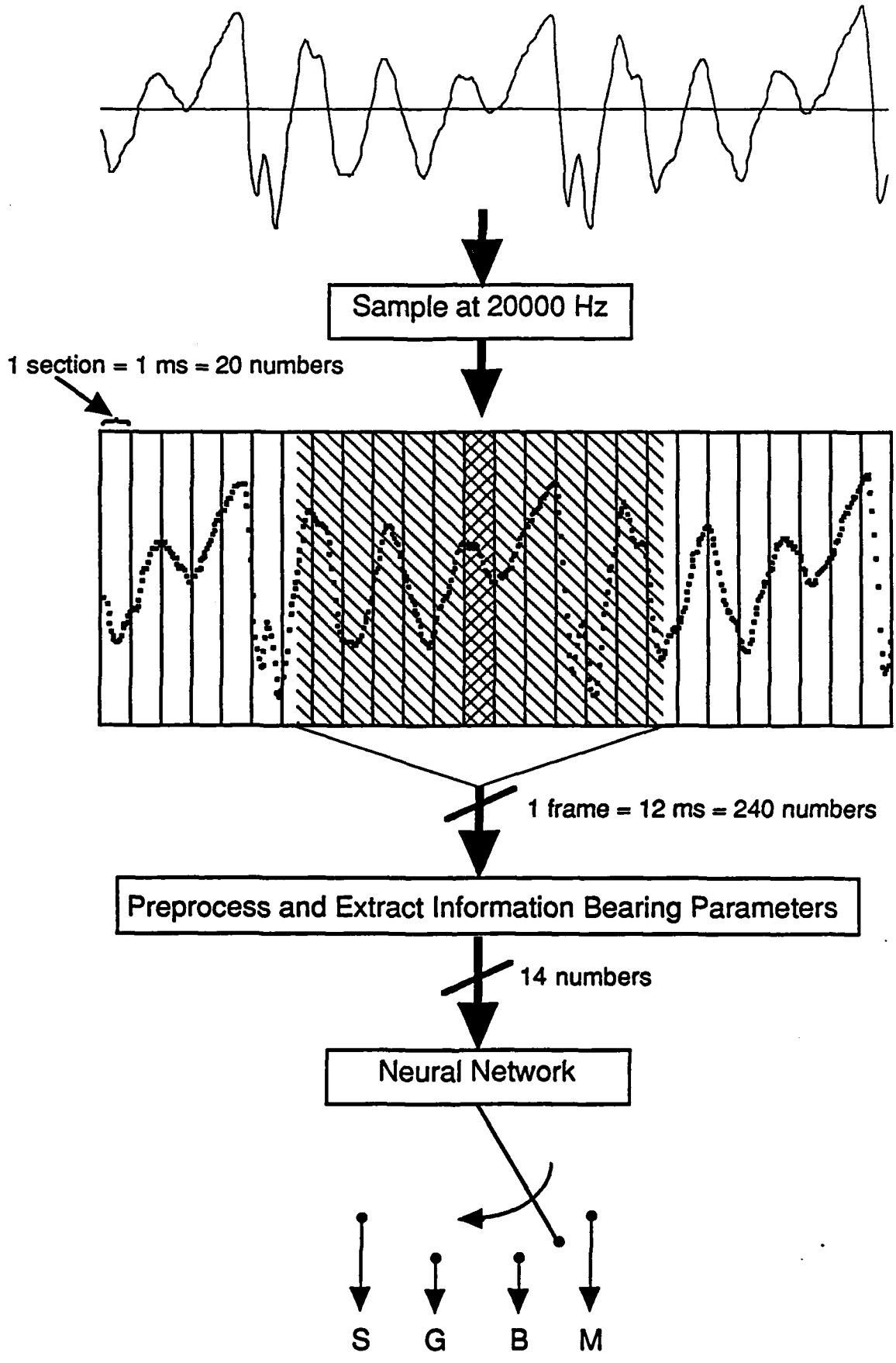


Figure 4-5: The classification problem from the neural network's perspective

of a typical speech waveform generated by the production of a glottal source sound is shown in the top panel of this figure. In the first operation, the speech signal is digitized at 20 kHz with 16 bit precision and stored on the computer. The signal is then treated as a set of contiguous 1 ms (20 sample) sections that are shown in the second panel as tall rectangles. A mapping is performed for each section. Because the pattern to be mapped has been digitized and quantized, there are 2.1410^{96} possible input patterns. The mapping is not performed on the raw output of the digitizer, but instead a set of 14 information bearing parameters are extracted and passed to the neural network. These parameters are extracted from more than one section of the input signal. Specifically, the parameters are computed from a 12 ms (240 sample) portion of the input signal centered on the section to be classified. This portion of the input signal is referred to as a frame. In the figure, the hatched portion of the signal symbolizes a frame and the double hatched portion represents the section to be classified. As shown by the four position switch in the bottom of the figure, the range of the mapping consists of the four categories (S/G/B/M) defined by the location of the sound sources in the human vocal tract. That mapping that the neural network performs generates a label for the center section in the frame. The frame is advanced 1 section (1 ms or 20 samples) forward and the process is repeated.

4.3.3 Neural Network Architecture

Neural networks are really a family of models, that can be distinguished by the topology, the training rules, and the node characteristics. The general framework

that is used in this work is a strictly layered feedforward network in conjunction with the supervised learning procedure called backpropagation [98]. The particular details and the justification of this framework will now be described.

Supervised learning means that during training, labels that specify the correct class for each input pattern are used in adapting the weights. The other alternative is unsupervised learning. This latter approach is typified by the work of Kohonen [99] where clusters are automatically formed based on the input patterns. Success has been shown for tasks involving speech recognition for both unsupervised [100] and supervised [101, 102, 103] learning paradigms. For the S/G/B/M classification task, I believe that supervised learning is a more straightforward approach.

Before the node characteristics and the details of the learning algorithm can be described, some basic terms need to be defined. Let o_{pj} be the output of unit j upon presentation of pattern p, let t_{pj} be the target output for the jth unit for pattern p, let w_{ij} be the weight from unit i to unit j, let θ_j be the bias for unit j, let η be the learning rate, and let α be the momentum rate.

The activation function that is used is the standard sigmoid or logistic (any function that is differentiable and nondecreasing will suffice) defined as

$$o_{pj} = \frac{1}{1 + e^{-(\sum_i w_{ij} o_{pi} + \theta_j)}} \tag{4.35}$$

The power of this multi-layer network is due to the nonlinear nature of the activation function. A function f is defined to be linear if and only if both of the following

equations hold

$$f(kx) = kf(x) \tag{4.36}$$

$$f(x_1 + x_2) = f(x_1) + f(x_2) \tag{4.37}$$

where k is any real number. Since neither of these equations are satisfied for the sigmoidal function, it is clearly non-linear. It has been shown [104] that if the activation function is linear, then for every multi-layer network there is an equivalent single-layer network. This is a severe limitation, since single-layer networks can only solve problems that are linearly separable. A common example of a problem that is not-linearly separable and hence one that can not be handled by a single-layer network or a multi-layer network with a linear activation function is the computation of a boolean exclusive OR.

The backpropagation learning procedure is based on the generalized delta rule. During the training phase, the weights are updated after each pattern is presented according to the following equation

$$\Delta w_{ij}(n+1) = \eta \delta_{pj} o_{pi} + \alpha \Delta w_{ij}(n) \tag{4.38}$$

where

$$\delta_{pj} = \begin{cases} o_{pj}(1 - o_{pj})(t_{pj} - o_{pj}) & \text{for output units} \\ o_{pj}(1 - o_{pj}) \sum_k \delta_{pk} w_{jk} & \text{for hidden units} \end{cases} \tag{4.39}$$

The calculation for the error is as follows

$$E_p = 1/2 \sum_j (t_{pj} - o_{pj})^2 \tag{4.10}$$

The learning procedure is a gradient descent approach towards minimizing this error function.

As stated before, the topology of the network that I am using is one that is strictly layered, feedforward, and the successive layers are completely connected (see Figure 4-6). By definition, a feedforward network does not contain any directed cycles, whereas a feedback or recurrent network [105] does. The simpler feedforward topology was used since Minsky and Papert have shown [106] that for every recurrent network there exists a feedforward network that behaves identically.

Although the original backpropagation algorithm was formulated only for strictly layered networks, Kimura has provided [107] a direct algebraic proof that the back-propagation learning rule is applicable to an arbitrary acyclic network. Therefore, the strictly layered approach was not chosen by necessity, but because it was postulated that there was no explicit need for lateral connections or connections that bypass layers.

Because of the previously discussed limitation for single-layer networks, at least one hidden layer is necessary for a task as complex as the S/G/B/M classification. By definition, a hidden layer is a layer of nodes that is not directly connected to either the input or the output. It can be shown [88] that a perceptron with a single hidden layer can only form decision regions that are convex. Although there was no direct evidence that complicated decision regions were needed, the implementation I have chosen is one with two hidden layers.

Since there are 14 information bearing parameters, the neural network has 14

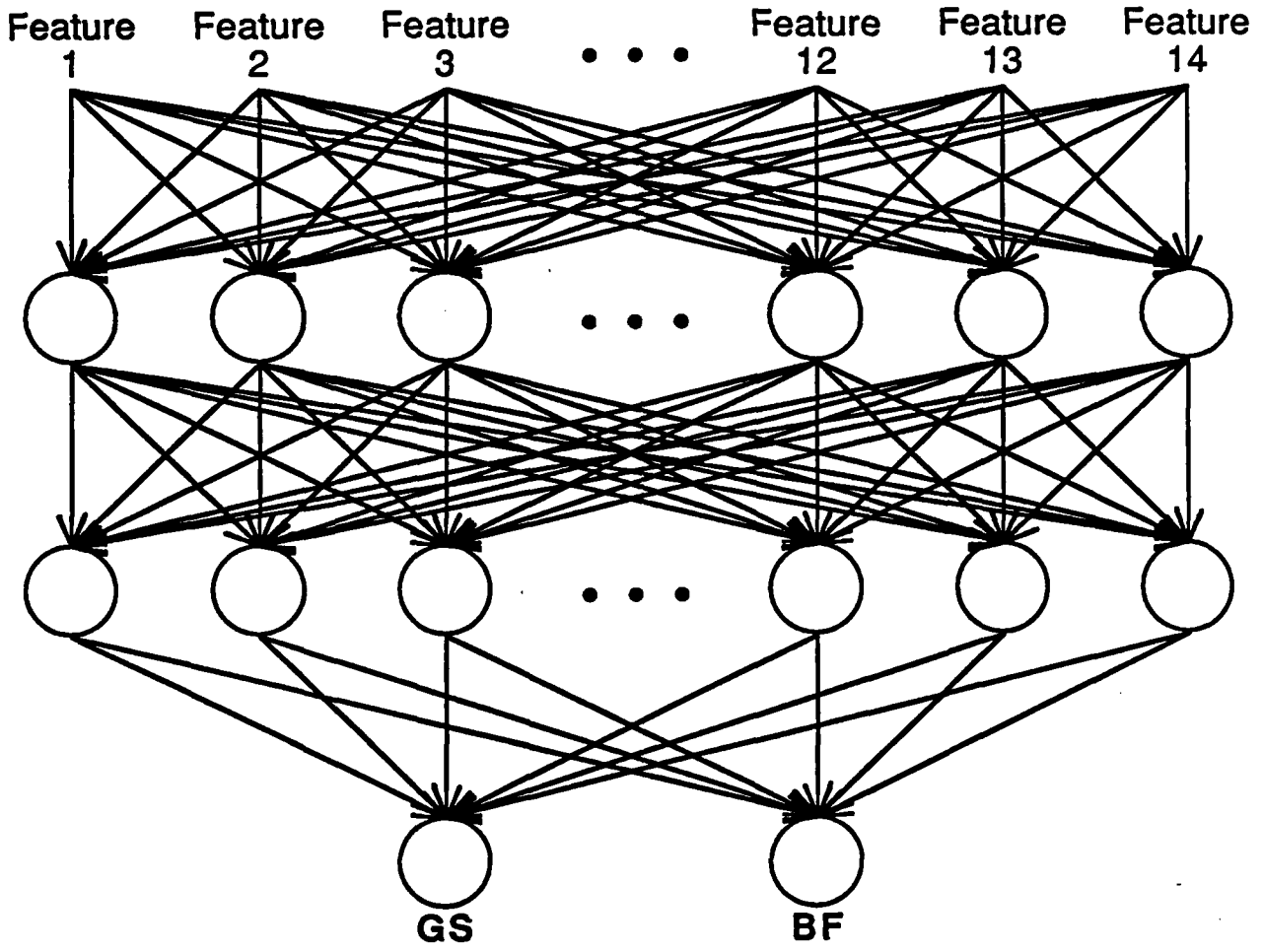


Figure 4-6: Topology of the Neural Network.

continuous-valued inputs. No contextual information is used: all of the input features are from the same frame. The exact number of nodes needed in the hidden layers was determined empirically. Although the initial estimate was roughly ten in each layer, the most successful classifier was built with 14 units in the first hidden layer and six units in the second hidden layer.

There are two output units, referred to as GS and BF. If the activation value of an output unit is ≥ 0.5 then the output unit is considered on; otherwise it is considered off. The classification label can be obtained by applying the mapping specified in table 4.1. Another alternative that was considered for the output coding was the one used by Gorman and Sejnowski [108]. In their system, which was designed to distinguish between the sonar signal returned from a rock and the signal returned from a metal cylinder, there is one output unit for each possible category. Using this scheme, four outputs would be required for the S/G/B/M classification. The output coding scheme of having one output unit for each possible sound source location, as opposed to one output unit for each possible class, has been used because it is a better representation of the underlying problem.

Table 4.1: Neural network output to real world label mapping.

GS	BF	Label
off	off	Silence
off	on	Burst Friction
on	off	Glottal Source
on	on	Mixed

The initial weights of the network are set to small random numbers uniformly distributed between -0.95 and 0.95. All of the biases are set to 0.5. The learning rate

is set to $\eta = 0.05$ and the momentum term is set to $\alpha = 0.025$.

Frequently, each time a neural network is trained with different starting states (random weights) different internal solutions (pattern of weights) are obtained. More often than not, these individual networks yield systems that exhibit different performance for both the training set and for generalization capabilities. Because of this sensitivity to the initial conditions, many different networks need to be trained in order to develop the range of expected performance. For all experiments involved in this study, unless explicitly noted, at least four different networks were trained to solve the task. It is important to note that the results reported for specific experiments are not for the average case, but represent the performance of the best network.

4.3.4 Training set presentation

Although it is useful to have networks with faster convergence times, the emphasis for optimizing the training set presentation was based on the premise of improving the percent correct performance of the classifier.

On each epoch (a complete pass through the training data) the order of the presentation of patterns is randomized. Also, the "learned" patterns are presented less frequently. This is motivated by two concerns. An intuitive example taken from Kimura [109] explains the first. When a person is developing his vocabulary, no one would suggest that he make 100 complete passes through the dictionary. The words that he has learned need not be presented on every pass through the dictionary. Instead he would generally focus his efforts (or training) on words which he has not

learned. Secondly, this method can be used to to minimize the over/under representation of particular input classes. For example, in the training set 6.6% of the segments are mixed sounds and 57.2% are glottal source sounds. If it takes an equivalent number of cycles to learn a particular class, then many extra training cycles for the over represented class would need to be performed, while the network is trying to master the under represented class.

Three different methods were evaluated for implementing this concept. All of these methods operate by altering the probability that a pattern is presented on any given epoch, based on the correctness of the previous presentation of that pattern.

The first and simplest method is to always present the pattern if the previous answer for that pattern is incorrect. Alternatively, if the previous answer is correct, the pattern would not be presented for the next n epochs. This method was tested for values of n between 2 and 5 and it did not significantly improve the performance (percent correct), although convergence times were reduced by roughly 15 percent when $n = 4$.

Caution must be exhibited when calculating convergence "time", since the standard unit of measurement (cycles or epochs) changes when modifications are made to the learning algorithm, For example, if a new algorithm reduces the number of cycles needed for convergence by 50 percent, but doubles the computational complexity of each cycle, then the new algorithm will take an identical amount of time to converge. Therefore, unless explicitly noted, whenever convergence times are compared, the comparison is based on CPU seconds.

Although the next method was developed independently, it is basically a modification of the focused-attention backpropagation algorithm presented by Hoskins [110]. In Hoskins' method, a random pattern p is presented with a probability based on the error term the last time it was presented:

$$P(\text{presenting pattern } p) = \begin{cases} \sigma & E_p < \text{pattern error tolerance} \\ 1 & \text{otherwise} \end{cases} \tag{4.41}$$

where σ is a fixed probability less than one. The algorithm that I developed calculates the probability in the following fashion:

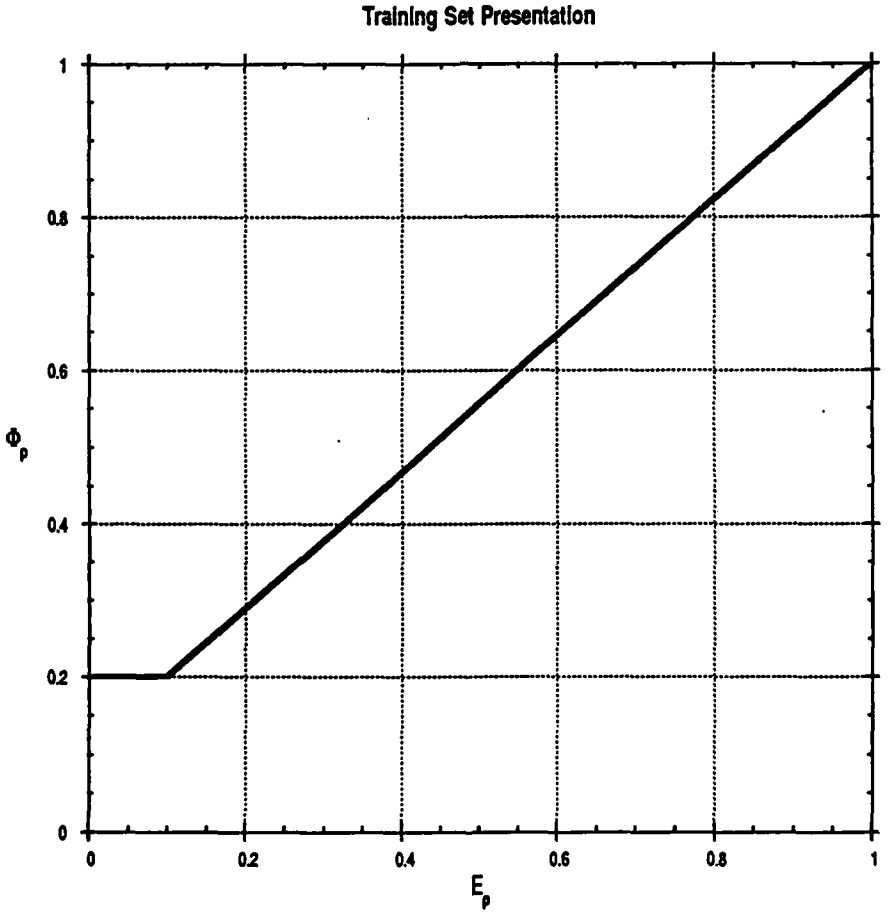
$$P(\text{presenting pattern } p) = \begin{cases} \sigma & (E_p < \text{pattern error tolerance}) \ \& \ (\text{correct response}) \\ 1 & \text{otherwise} \end{cases} \tag{4.42}$$

Like the results reported by Hoskins, this method resulted in faster (roughly 20 percent for $\sigma = 0.3$ and an error tolerance of 0.12) learning. Unfortunately, the percent correct was not an improvement over standard backpropagation.

The method that showed the best success will now be presented. In the most general form, a random pattern, p, is presented with a probability based on the error the last time it was presented

$$P(\text{presenting pattern } p) = \Psi(E_p) \tag{4.43}$$

After much study, an empirical function was developed for Ψ that significantly improved not only the convergence time, but also the percent correct performance. The function for Ψ is presented in Figure 4-7. Compared to standard methods, this


Training Set Presentation
1
0.8
0.6
0.4
0.2
0
Φ_p
0
0.2
0.4
0.6
0.8
1
E_p

method improved the generalization capabilities roughly 2.0 percent. The convergence time was also improved by approximately 20 percent.

4.4 POSTPROCESSING

To improve the performance of the S/G/B/M classification, a postprocessing scheme has been developed to correct for obvious mistakes in the labeling process. For example if a 1 ms burst friction label is centered within a 100 ms glottal source section, then the label should be changed to glottal source. The only input to this stage is the stream of classifications obtained from the neural network. This is the only stage of the system where information over multiple frames is used.

In the postprocessing stage, a conservative approach is taken because of the desire not to mistakenly change any correct labels. The purpose of postprocessing is error correction; the burden of classification is left to the neural network. Therefore, statistically the desired loss function is such that more than one error must be corrected for every accurate label that is incorrectly changed. The philosophy is similar to our legal system; a frame is considered correct unless it can be proven beyond a reasonable doubt that a mistake has been made.

Since most practical usage dictates that extra segments are preferable to having too few, I have tried to err on the side of having "noisy" data as opposed to overly "smoothing" the data.

One possible way to detect and correct obvious errors in the stream of classifications is to develop a small set of rules. An examination of the training set shows that,

as expected from the mechanics of speech production, all possible pairwise segments do occur. More formally, if the stream of labels is viewed as a grammar, then the grammar is unrestricted with respect to the four labels. Hence, simple rules of the form "mixed can not follow silence" are not applicable.

Most filtering algorithms expect the input data to be either continuous or ordered. Since the four classes for this study are clearly not ordered (or continuous), the calculation of a conventional average is not a useful concept.

The first approach that was investigated was to use a majority vote over a small fixed number of frames (window). Class membership is examined over a fixed number of frames and if there is a majority vote for any class then the center frame in the window is reassigned to the class with the largest vote. This method was developed in an effort to adapt the concept of a non-linear median smoother [111] to non-ordered discrete data. This approach was tested for window sizes of 3, 5, 7, and 9 frames (classifications). The best performance increment (compared to no postprocessing) was approximately 0.5 percent with a window size of 5 frames.

The best approach turned out to be one that used a set of rules based on the minimum expected duration of each of the four classes in the training set. A segment is defined as a contiguous set of frames all belonging to the same class, bounded on each side by a frame from a different category. Moving forward in time, the length of each segment is checked against a minimum segment length for each category. The shortest segment lengths allowed for silence, glottal source, burst friction and mixed categories are 5 ms, 6 ms, 3 ms, and 6 ms respectively. If a segment is found with

a length less than its allotted duration, then the whole segment is reassigned to the category of the previously (in time) encountered segment.

5. IMPLEMENTATION

The computer that has been used for most of this research is a VMS VAXstation 3200. The vendor (Digital Equipment Corporation) rates the relative performance of this machine as 2.7 times the performance of a VAX 11/780 (VAX MIPS).

Although the number of lines of code is not always a meaningful metric, it does allow one to make an estimate of the effort involved in a software development project [112, chapter 8]. For this research, over 12400 lines of code have been written. This development has taken the better part of two years. It should also be mentioned, that miscellaneous portions of this study have also used software previously developed at CID and various commercial software packages.

Next, the major software components that were engineered for this study will be listed, along with a brief description of each.

5.1 SAMPLING AND DISPLAYING A SPEECH WAVEFORM

To control analog-to-digital (A/D) and digital-to-analog (D/A) operations, as well as to perform simple editing and windowing operations upon sampled waveforms, I have developed a package referred to as SINS, which is an acronym for Speech IN the auditory perceptual Space. SINS has been tested with as many as 16 users logged onto the system at once, indicating that it is feasible to do many real-time operations on a multi-user computer running the VMS operating system.

The SINS package was an important component for two portions of this work. First, SINS was used to digitize the entire speech database. More importantly, it was the primary tool that was used for performing the manual classification. The extreme ease of use and flexibility of this package provides the speech scientist with a powerful tool. Specifically, the ability to quickly manipulate and examine waveforms in detail greatly facilited the manual classification.

This interactive waveform editor is designed to work with a DigiSound-16 system connected to a VMS VAX with a Q-BUS using a custom interface developed at CID [113], along with a DRV11-WA. SINS is used to digitize our audio tapes recorded on a JVC VCR in the anechoic chamber. SINS is capable of reading and writing many different file formats including files which are compatible with ILS, a popular signal processing package that is commercially available and in use at CID. The file level compatibility with ILS allows the flexibility of using the more than 100 functions provided by ILS. This was useful in the initial stages of this project for developing and choosing the set of information bearing parameters.

Written in a modular fashion, the SINS software alone is roughly 4900 lines of FORTRAN code. The DigiSound-16 is never accessed directly, since all input and output for the DigiSound-16 system is performed through the DigiSound-16 library which I have also developed. The DigiSound-16 Library is roughly 500 lines of FORTRAN code. There are only three SINS commands that call routines from the DigiSound-16 library: play, record, and set the sampling rate. To enable this package to work with a a different D/A and A/D system would simply require a rewrite of these three

routines.

SINS provides a graphical interface for a Tektronix 4010 compatible terminal. All graphics operations are implemented using the PLOT10 library (described in the next section), allowing this software to be used upon any type of terminal supported by the PLOT10 library. To enable the graphics to work with a different type of terminal, would simply require modifying the PLOT10 package. All other screen I/O (user input and prompting) is performed using the standard DEC Screen Management routines (SMG Run Time Library).

5.2 GRAPHICS

To simplify and standardize the writing of software which utilizes graphics, I have developed a set of 2-dimensional graphics subroutines and compiled these into a library which I call PLOT10. This library provides a functionally complete graphical interface to any device that can emulate the Tektronix 4010 series of terminals, The PLOT10 library is roughly 3900 lines of FORTRAN code.

This library has enabled us to develop many applications that can display graphic output. It has allowed researchers whose only familiarity with computers is FORTRAN to develop graphical software without involving them in the details of sending escape sequences and cryptic address coordinates.

To handle the various peculiarities of different Tektronix 4010 emulations at run-time (as opposed to compile or link time), this graphics package utilizes a system wide text file that describes the individual characteristics of the particular terminal

type being used. For example, in this file attributes such as terminal resolution and escape sequences for entering and exiting graphics mode are stored. This frees the programmer from dealing with the intricacies of each particular terminal, providing some degree of device independence. This package works on all of the graphics terminals that we have access to including DEC VT240s, MicroTerm Ergo-301s, Graphon GQ-140s, and HP2623s. Hardcopy can either be obtained by screen dumps from any of our HP2623s or we can direct the graphics package to use our LN03 laser printer for the output. These routines were meant to be called from FORTRAN, but if the proper calling conventions are maintained, they may be called from any other language.

5.3 PREPROCESSING AND FEATURE EXTRACTION

The preprocessing and extraction of the information bearing parameters is all done by a single 1500 line FORTRAN program. Modularity and independence of computations had the highest priority when this program was designed. The cost of this extreme modularity was computational efficiency. For example, the routine that calculates zero crossing rate and the routine that calculates the reversal rate could have been combined into one routine and by doing so reducing the execution time for these two routines by almost a factor of two. Of course, the advantage is that the individual calculations are much easier to modify and maintain. Since this research project was experimental in nature, the individual routines were constantly being modified in search of a better classifier.

Most of the preprocessing parameters can be changed at run time, using command line options in the standard VMS fashion. Some of the preprocessing steps that can be performed by this program are notch filtering (default of 60 Hz), highpass filtering (default 50 Hz), custom filtering (by specifying a file with the time domain coefficients), preemphasis (default parameter 0.98), and Kaiser-Bessel windowing, After the preprocessing has been performed, the information bearing parameters are extracted.

Ideally, in a production system, the information bearing parameters would be computed in parallel in real time. This would allow the whole system to be much closer to a physiological implementation. This type of scheme has not only been proposed, but in fact, Mead [114] has designed both an analog silicon retina and an analog electronic cochlea in an attempt to build dedicated physiologically based pre-processors modeled after the human sensory systems. Unfortunately, until these types of devices are generally available, sequential software simulations will have to suffice.

After the information bearing parameters have been extracted an output text file is created. Each line in this file corresponds to one frame. Every line has 15 space delimited fields corresponding to the frame number and the 14 information bearing parameters. The information bearing parameters are represented with four digit accuracy. Because easy manipulation of the data was required, a text file as opposed to a binary data file, is used. The cost of this flexibility is 105 bytes per frame.

5.4 NEURAL NETWORK SIMULATOR

As was indicated earlier, a connectionist approach is used for the class assignment stage. There were a couple of existing connectionist simulators that I considered using for this project. Unfortunately, none were suitable.

The one that seemed to have the most potential for this project, the Rochester Connectionist Simulator [115], had several problems. Most notably the commands to save and restore the state of the network do not work with backpropagation. Although the documentation is lengthy, it is lacking; whenever details are needed, the source code must be examined. The code also has some disturbing flaws; specifically the sigmoidal activation function does not handle floating point overflow, and can cause the program to crash. Also, the simulator is UNIX specific, and would be fairly complicated to port to VMS. Additionally, because each unit and each link are simulated with function calls, the simulator is slower than a dedicated backpropagation simulator.

Therefore, I decided the best solution was to write my own simulator. For this application, the C programming language is clearly much better than FORTRAN. This outweighed the fact that all earlier stages of this project were written in FORTRAN.

The general purpose backpropagation simulator that was developed consists of 1500 lines of portable C. The main features of the simulator are that it is fast, built on a foundation that is easy to modify, and it is portable. The simulator was designed to handle strictly layered feedforward networks with the backpropagation learning algorithm. The number of layers and the number of units per layer are specified

interactively at run time.

The simulator reads in the data files created from the information bearing parameter extraction stage, along with a file containing the manually labeled classifications. The simulator has an interactive command interface to control its operation and to examine the network. When the simulator is tested on a data set, it can create an output text file with the labels that the network has computed.

Because it takes such a long time to train the neural network, the time for preprocessing, extraction of information bearing parameters, and postprocessing is negligible. Hence, the effort for performance optimization was focused on the backpropagation simulator.

The computational bottleneck in the learning algorithm is the exponential function that is used for the sigmoidal activation function. Although this is hardware and software specific, I have performed a timing analysis for floating point computations on the VMS VAXstation 3200 that was used for this project. The standard exponential operation takes 250 microseconds. In comparison, multiplication takes 1.4 microseconds, addition takes 1.4 microseconds, and an assignment operation takes 0.65 microseconds. To remove the bottleneck caused by the exponential function, a table lookup function was implemented that reduces exponentiation to 16 microseconds. The implementation has 5000 entries and takes advantage of the identity

$$e^{-x} = \frac{1}{e^x} \tag{5.1}$$

Although not carefully evaluated, it was informally observed that networks that used the table lookup converged in less cycles than networks without.

The best classifier takes roughly 38 CPU hours to train using the neural network simulator.

5.5 POSTPROCESSING

In this stage, only very simple computations are performed. The postprocessing is done by a single 100 line portable C program. The basic function of this program is to read in a text file of classifications that was created by the connectionist simulator and to use a simple set of rules to correct any obvious errors. The output is also sent to a text file.

6. MEASUREMENTS, RESULTS, AND DISCUSSION

First some measurements describing the training sets from both subjects (JH and JW) will be presented. Next, different methods will be discussed for judging the performance of the classifier followed by an examination of the accuracy of the classifier. Finally, an analysis and discussion is presented for the results, for the neural network, and for postprocessing.

At this point, it is useful to remember that the portion of the data set that is referred to as training is the first paragraph of the rainbow passage while the rest of the passage is referred to as the testing set. When neither training nor testing is mentioned, the assumption is that the entire data set is being used. Also, the two data sets that correspond to the two different speakers will be referred to as JH and JW.

6.1 MEASUREMENTS

For each training set, an analysis has been performed to describe the hand labeled classifications. From these measurements, a better understanding can be obtained of the frequency of occurrence of each of the four categories and of the duration of individual segments.

The number of frames in each category is presented in table 6.1. Since both subjects recited the same passage, it is reasonable to expect similar class compositions

and this expectation is reflected in the data. One item to be noted, is that each class is not equally represented in the training set. This non-equal class representation is important for training the classifier and is addressed in the training set presentation procedure described in section 4.3.4. Also, from this data, it can be seen that a classifier that simply labels all sounds as glottal source would be accurate at least 55.9 percent of the time.

Table 6.1: Class Composition for the Training Sets

Data set	S	G	B	M
JH	7958 ms	17834 ms	3317 ms	2066 ms
	25.5%	57.2%	10.6%	6.6%
JW	7101 ms	19004 ms	5521 ms	2374 ms
	20.9%	55.9%	16.2%	7.0%

Next, the number of segments for each category is presented. Remember, a segment is a contiguous set of identical classifications bounded on each side by a classification from a different category. Although, the data in table 6.2 was not used for designing the classifier, it provides a breakdown that is interesting from a speech production perspective.

Table 6.2: Number of segments in the training sets

Data set	S	G	B	M
JH	62	107	68	80
JW	57	103	82	122

A statistical description of the durations of individual segments will now be presented. The analysis for the training set from speaker JH (male) and for the training set from the speaker JW (female) are presented in tables 6.3 and 6.4 respectively. The distributions from each of the categories (S, G, B, M) are all highly skewed (more

segments with shorter durations) for both speakers. One important item to note in these tables, is that mixed segments are the shortest and glottal source segments are the longest. The data in this table are in agreement with expectations from the following simple linguistic analysis of speech production. If we assume that a syllable is composed of one glottal source segment, one burst friction segment, and one mixed segment, then the data in these tables is in line with the rule of thumb duration of 250 ms for a syllable in normal conversational speech.

Table 6.3: Durations of each segment for speaker JH

Statistic	S	G	B	M
Mean	128 ms	167 ms	49 ms	26 ms
sd	165 ms	133 ms	43 ms	21 ms
Min	4 ms	5 ms	2 ms	5 ms
Max	619 ms	709 ms	177 ms	87 ms

Table 6.4: Durations of each segment for speaker JW

Statistic	S	G	B	M
Mean	125 ms	185 ms	67 ms	19 ms
sd	257 ms	158 ms	55 ms	19 ms
Min	2 ms	5 ms	2 ms	1 ms
Max	1252 ms	866 ms	232 ms	139 ms

Also, this data proved useful in the development of the postprocessing algorithms. From this data, minimum allowable segment lengths were constructed for correcting mislabelings from the connectionist network.

Next, some statistics are presented on each of the 14 information bearing parameters. The means and standard deviations for each class are presented in tables 6.5 and 6.6 for the JH and JW training sets. Although these statistics were not used in the development of the classifier, they are useful for characterizing individual information

bearing parameters. It can be observed, that for many of the information bearing parameters the four classes show significant overlap. At first examination, one might expect that this overlap severly limits the ability of the classifier and implies that the feature set is not appropriate for the classifier. This overlap was not only expected, but it was one of the motivating factors for not using a statistical classifier.

Table 6.5: Means and standard deviations for the four classes using the JH training data

	S	G	B	M
	mean (sd)	mean (sd)	mean (sd)	mean (sd)
RMS	0.130 (0.064)	0.427 (0.096)	0.406 (0.085)	0.346 (0.098)
ZC	0.969 (0.370)	0.388 (0.138)	1.360 (0.599)	0.892 (0.528)
LPCR	-1.753 (0.183)	-0.875 (0.349)	-1.509 (0.206)	-1.504 (0.292)
AC	0.076 (0.411)	0.768 (0.134)	0.092 (0.403)	0.272 (0.382)
RR	0.938 (0.278)	0.377 (0.122)	0.929 (0.280)	0.772 (0.260)
CEP	0.319 (0.212)	1.884 (1.857)	0.373 (0.264)	0.470 (0.383)
S1	0.340 (0.348)	0.777 (0.272)	0.186 (0.307)	0.411 (0.395)
S2	0.626 (0.325)	0.911 (0.141)	0.567 (0.312)	0.634 (0.328)
S3	0.630 (0.312)	0.822 (0.200)	0.733 (0.263)	0.737 (0.266)
S4	0.632 (0.304)	0.644 (0.233)	0.763 (0.263)	0.718 (0.279)
S5	0.612 (0.312)	0.360 (0.217)	0.671 (0.283)	0.566 (0.319)
S6	0.569 (0.336)	0.167 (0.188)	0.624 (0.304)	0.475 (0.344)
S7	0.582 (0.339)	0.183 (0.188)	0.650 (0.316)	0.519 (0.336)
S8	0.565 (0.352)	0.186 (0.193)	0.587 (0.326)	0.482 (0.332)

In fact, neural networks can solve many problems that are not easily solved by statistical approaches. A good example of a problem that most statistical approaches fail to solve is the two-spiral problem [116]. The goal in this problem is to distinguish between points on two intertwined spirals in a unit square. This is illustrated in Figure 6-1 with one hollow spiral and one solid spiral. Using two input features, the X and Y coordinates, a two layer backpropagation network was able to properly distinguish points on the two spirals. It can be clearly seen, that most statistical

Table 6.6: Means and standard deviations for the four classes using the JW training set

	S	G	B	M
	mean (sd)	mean (sd)	mean (sd)	mean (sd)
RMS	0.138 (0.070)	0.413 (0.091)	0.440 (0.116)	0.336 (0.103)
ZC	0.944 (0.405)	0.343 (0.162)	1.575 (0.783)	1.043 (0.673)
LPCR	-1.738 (0.248)	-1.142 (0.350)	-1.370 (0.366)	-1.555 (0.289)
AC	0.111 (0.398)	0.745 (0.211)	0.008 (0.453)	0.124 (0.452)
RR	0.911 (0.266)	0.409 (0.179)	0.991 (0.315)	0.894 (0.307)
CEP	0.309 (0.234)	2.386 (2.241)	0.305 (0.109)	0.459 (0.367)
S1	0.336 (0.338)	0.792 (0.285)	0.185 (0.288)	0.352 (0.367)
S2	0.642 (0.326)	0.868 (0.187)	0.496 (0.338)	0.575 (0.342)
S3	0.647 (0.319)	0.730 (0.252)	0.625 (0.319)	0.639 (0.320)
S4	0.621 (0.307)	0.563 (0.290)	0.716 (0.275)	0.666 (0.298)
S5	0.632 (0.305)	0.417 (0.270)	0.703 (0.270)	0.619 (0.310)
S6	0.584 (0.325)	0.250 (0.222)	0.668 (0.297)	0.566 (0.326)
S7	0.570 (0.347)	0.242 (0.227)	0.664 (0.314)	0.584 (0.334)
S8	0.511 (0.352)	0.188 (0.216)	0.551 (0.343)	0.476 (0.348)

approaches would fail for this problem.

In this study, the basic premise is that the neural network is not dependent upon the statistical separation of the 14 information bearing parameters. By using structural relationships in the input features, the network seems to perform successful classifications.

6.2 PERFORMANCE CRITERIA

There are many methods that I have investigated for judging the accuracy of the classifier. Two of these are used to report the results of the system.

The most straight forward method for measuring percent correct is to simply count the number of disagreements. This simple scheme is detailed by the error matrix presented in table 6.7. Each entry in the table corresponds to the error associated with

a particular classification. The total percentage agreement is calculated by summing the errors for each classification and dividing by the total number of classifications. When results are presented, this method is referred to as SIMPLE. Because of the nature of the classification task, this is not necessarily the most meaningful method.

Table 6.7: This simplest error weighting scheme (SIMPLE)

	S	G	B	M
S	0	1	1	1
G	1	0	1	1
B	1	1	0	1
M	1	1	1	0

Another metric that was used for rating percent agreement, is based on a Hamming distance, which is defined as the number of features that differ in a pattern. The S/G/B/M classification can be thought of as having two features corresponding to the two different possible locations for sound source generation. In this scheme, each classification is viewed as two decisions and 1/2 credit is scored for each correct decision. This partial credit scheme is detailed by the error matrix presented in table 6.8. Once again, the percentage agreement is obtained by summing the appropriate entries in the table and dividing by the total number of classifications. Whenever results are presented, the label HAMMING refers this method. It can be seen that this method will always yield higher percentage agreements than SIMPLE. The attractive aspect of this metric is that it closely reflects the underlying problem.

This approach can be generalized by using an arbitrary error matrix. The two methods listed above (SIMPLE and HAMMING) are based upon the nature of the classification and were not designed with any particular application in mind. If one

Table 6.8: Error weighting scheme based on Hamming distance (HAMMING)

	S	G	B	M
S	0	0.5	0.5	1.0
G	0.5	0.0	1.0	0.5
B	0.5	1.0	0.0	0.5
M	1	0.5	0.5	0.0

had a particular application in mind another matrix might be more desirable. The entries in the matrix might reflect the estimated costs of making particular mistakes. For example, if this system was used as a silence detector then the error matrix presented in table 6.9. would be most appropriate.

Table 6.9: Error weighting scheme for silence detection

	S	G	B	M
S	0	1	1	1
G	1	0	0	0
B	1	0	0	0
M	1	0	0	0

Fundamentally, there is some inaccuracy introduced by the manual classification because the "correct" classification is not always obvious. Therefore, a better method of interpreting accuracy is percent agreement as opposed to percent correct. Unless otherwise noted, all results presented in this work are based upon precent agreement with the author's manual classification.

Because of the inherent ambiguity of the classification, it is useful to examine the percent agreement between speech scientists. Even for classification schemes similar to this, quantitative results of percent agreement between speech scientists have not been reported. Therefore, I have tried to obtain an estimate of the theoretical maximum percent agreement. For this estimate, three scientists classified the first two sentences

from the recording of the male (JH) reciting the rainbow passage:

> "The rainbow passage. When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow."

The recorded signal is 6.9 seconds long with roughly 1.2 seconds of silence surrounding the two sentences.

The percentage of classifications with all three scientists in agreement was 87.8. Whereas, in 98.7 percent of the frames at least two out of three scientists were in agreement. Pairwise agreements are presented in table 6.10.

Table 6.10: Pairwise agreement over 6.900 seconds of signal

Scientist	Scientist	SIMPLE	HAMMING
SJS	MSF	93.1	96.2
SJS	JDM	91.7	95.7
MSF	JDM	89.6	94.6

It can be seen from this table that when the SIMPLE and HAMMING methods are used the percent agreement between scientists is roughly 91.5 percent and 95.5 percent respectively. To verify this preliminary data, a longer passage (39.450 seconds) was manually classified by two scientists (SJS and MSF). The percent agreements were 91.5 and 95.0 for the SIMPLE and HAMMING methods. This extended comparison yields results similar to the shorter (6.9 second) comparison, indicating that these figures are reasonable upper bounds for an optimal automatic classification system.

A more detailed description of the mistakes that are made is presented in table 6.11. Each entry in the table corresponds to the percentage correct for a particular confusion. For example, 1.1 percent of the frames that SJS classified as glottal source

were classified as silent frames by MSF. It should be noted that silence is identified the most consistently and that mixed classifications are frequently confused with glottal source and burst friction sections. It will be shown, that these trends are also exhibited by the mistakes made by the neural network.

Table 6.11: Confusions between speech scientists for 39.450 seconds of signal

	SJS identified as			
	S	G	B	M
MSF identified as				
S	98.2	1.1	22.2	3.1
G	0.4	96.2	0.6	20.4
B	0.2	0.4	72.9	10.9
M	1.2	2.3	4.4	65.7

6.3 RESULTS

In total, the results from twelve experiments are presented in this section. These twelve are based on the manipulation of three orthogonal ($12 = 2 \times 3 \times 2$) conditions.

The first variable was the training set. Based on the two different speakers used in this study, two different networks have been developed as potential classifiers. The first one was developed using the JH training data and the other was created with the JW training data. Performance results are presented separately for both of these networks. For conciseness, the first network is referred to as NETJH and the second network is referred to as NETJW.

The next factor involved the data set that the network was tested with. The results are presented separately for each of the three possible data sets. These are the data set that was used for training, the test set for the same speaker, and finally

the entire data set from the other speaker. This condition allows the generalization capabilities of the network to be investigated.

The last condition that was examined was whether or not the postprocessing stage was used. If postprocessing is performed the result is referred to as POST; otherwise the result is referred to as RAW. The manipulation of this condition allows the direct evaluation of the performance of the neural network without the additional error correction stage.

The results of the twelve experiments involving two networks (NETJW, NETJH), optional postprocessing (RAW, POST), and the data set the network was tested on (train, same speaker, different speaker) will now be presented.

The overall performance of the NETJH network is presented in table 6.12. Each result from the six experiments with this network is presented using both metrics (SIMPLE and HAMMING) discussed in section 6.2 resulting in twelve entries in the table. Similarily, the overall performance of the NETJW network is presented in table 6.13.

Table 6.12: Performance for network NETJH

Data Set	RAW		POST	
	SIMPLE	HAMMING	SIMPLE	HAMMING
JH train	92.5	96.0	93.3	96.4
JH test	87.3	92.7	88.2	93.3
JW	89.0	94.6	89.4	94.7

Because the a priori probabilities for each class are not equal, performance has also been measured and analyzed on a per class basis. These data are presented in the form of confusion matrices. There are 12 tables corresponding to the twelve experiments

Table 6.13: Performance for network NETJW

Data Set	RAW		POST	
	SIMPLE	HAMMING	SIMPLE	HAMMING
JW train	90.8	95.7	91.1	95.9
JW test	90.0	94.7	90.5	95.1
JH	88.2	93.4	88.8	93.7

which involved two networks (NETJW, NETJH), optional postprocessing (RAW, POST), and different data sets that the network was tested on (train, same speaker, different speaker). Each entry is in the table is reported as a percentage correct in reference to the manual classification performed by the author.

The accuracy of the NETJH network is reported in six confusion matrices in tables 6.14 – 6.19. Similarly, the accuracy of the NETJW network is reported in six confusion matrices in tables 6.20 – 6.25.

Table 6.14: Confusions for NETJH RAW with JH training data

	Actual answer			
	S	G	B	M
NETJH identified as				
S	95.6	1.0	2.5	2.3
G	1.9	97.4	1.3	26.2
B	2.3	0.2	89.5	22.0
M	0.2	1.4	6.7	49.5

Table 6.15: Confusions for NETJH POST with JH training data

	Actual answer			
	S	G	B	M
NETJH identified as				
S	96.3	0.7	1.6	2.1
G	1.2	97.9	1.2	29.6
B	2.4	0.3	94.5	23.6
M	0.1	1.1	2.6	44.7

Table 6.16: Confusions for NETJH RAW with JH testing data

	Actual answer			
	S	G	B	M
NETJH identified as				
S	92.4	1.6	5.2	11.0
G	3.2	95.7	4.4	30.9
B	3.7	0.8	67.9	33.4
M	0.7	2.0	22.4	24.7

Table 6.17: Confusions for NETJH POST with JH testing data

	Actual answer			
	S	G	B	M
NETJH identified as				
S	92.8	1.2	4.9	9.2
G	2.5	96.6	3.4	33.8
B	4.1	0.9	73.8	37.6
M	0.7	1.2	17.9	19.4

Table 6.18: Confusions for NETJH RAW with JW data

	Actual answer			
	S	G	B	M
NETJH identified as				
S	96.2	1.0	9.6	18.3
G	1.3	95.9	3.3	16.7
B	2.0	0.7	83.3	43.5
M	0.5	2.4	3.8	21.5

Table 6.19: Confusions for NETJH POST with JW data

	Actual answer			
	S	G	B	M
NETJH identified as				
S	96.5	0.9	9.4	17.0
G	1.2	96.2	3.3	17.9
B	2.0	0.8	85.7	47.1
M	0.4	2.1	1.5	18.0

Table 6.20: Confusions for NETJW RAW with JW training data

	Actual answer			
	S	G	B	M
NETJW identified as				
S	93.9	0.7	3.6	13.5
G	2.1	97.8	0.5	23.1
B	3.9	0.3	92.0	40.4
M	0.1	1.1	3.9	23.0

Table 6.21: Confusions for NETJW POST with JW training data

	Actual answer			
	S	G	B	M
NETJW identified as				
S	94.4	0.6	3.2	14.0
G	1.8	98.2	0.3	24.4
B	3.8	0.3	94.4	44.3
M	0.0	0.9	2.2	17.4

Table 6.22: Confusions for NETJW RAW with JW test data

	Actual answer			
	S	G	B	M
NETJW identified as				
S	92.5	0.5	6.3	5.5
G	1.6	97.1	2.4	18.9
B	4.9	0.6	86.1	49.1
M	1.0	1.9	5.2	26.6

Table 6.23: Confusions for NETJW POST with JW test data

	Actual answer			
	S	G	B	M
NETJW identified as				
S	93.2	0.5	5.9	5.7
G	1.1	97.3	1.9	19.8
B	4.8	0.6	88.5	48.8
M	0.9	1.6	3.7	25.7

Table 6.24: Confusions for NETJW RAW with JH data

	Actual answer			
	S	G	B	M
NETJW identified as				
S	90.6	0.9	1.2	6.1
G	3.3	95.9	3.8	29.7
B	5.3	0.8	85.2	42.5
M	0.9	2.4	9.8	21.7

Table 6.25: Confusions for NETJW POST with JH data

	Actual answer			
	S	G	B	M
NETJW identified as				
S	91.2	0.7	1.1	5.5
G	3.0	96.7	3.7	32.5
B	5.3	0.8	87.6	44.6
M	0.5	1.7	7.6	17.4

6.4 ANALYSIS OF THE RESULTS

Many different experiments have been performed in an effort to develop a system that can successfully perform the S/G/B/M classification task.

As expected, when the errors are analyzed on a class by class basis, most of the mistakes occur for mixed sections. The effects of the poor performance for mixed sections is minimized because mixed segments are shorter in duration and because they occur less frequently.

A key question for any inductive learning machine is how well it handles new input data after the system has been trained. In fact, one could easily develop a table lookup scheme to memorize the input training data that could perform close to 100 percent. The problem with this scheme, is that when tested on other data sets performance would probably drop to around chance levels. Therefore, this network

was tested on other data not in the training set so that the generalization capabilities could be accurately assessed.

The network that was developed with the JH training data set seemed to master the classification quite well for the training data. In fact, the task was "learned" so well, that this network yielded a better percent agreement than would be expected if another scientist (besides for SJS) labeled the data. By itself, this does not imply that the generalization capabilities of the network would allow the network to successfully classify other data sets.

When this network was tested on data that it had not been exposed to during training, the performance dropped to roughly three to five percent below the expected accuracy of hand labeled material. These results imply that this network has probably overlearned the idiosyncrasies of the training set. But, the generalization capabilities of the network still allow the system to successfully handle novel signals.

The network that was trained with the JW training set also shows an expected performance degradation when tested on data that it has not been exposed to during training. This performance decrement is less than the decrement seen in the NETJH network.

Although the network that was trained with the JW training set did not learn the training set as well as the NETJH network learned its training set, a smaller decrement in performance was seen when tested on the novel speech. In fact the performance did not decline more than one percent when tested on speech from the same speaker and it dropped at most 2.5 percent when tested on speech from the

other speaker.

A curious item can be observed about the performance of the NETJH network. It performs better with data from the second speaker (JW) than it does when using the test set from the same speaker. Although the magnitude was not always as large, this result was consistently shown when different initial random weights were used for the network. This improved accuracy for a different speaker was not observed for the network that was trained with the JW training set. An examination of the confusion matrices indicates that the performance for glottal source sections is roughly the same for both data sets, the performance for mixed sections is a little better for the JH testing set, and that the performance for silence and burst friction sections is better for the JW data set. Since mixed segments account for at most 7 percent of the data, the contribution of the slight increment in accuracy for mixed sections for the JH testing set can be ignored. Hence, the reason that the NETJH network performs better with the data from the other speaker can be attributed to better performance on silence and burst friction sections from the other speaker. The most likely explanation is that the NETJH network has developed some internal representation based on the training data that allows it to more accurately classify silence and burst friction sections from the speech of the other speaker.

The best classifier was obtained with the network that was trained on the JW training data set. The accuracy of this classifier on data from the same speaker that was not used for training was 90.0 percent. It also achieved 88.2 percent accuracy when tested on speech from a novel (JH) speaker. When compared to the percent

agreement (91.5) between two speech scientists these results are quite impressive.

6.5 EXAMINATION OF THE NEURAL NETWORK

Although a neural network is an integral part of this system, the development and evaluation of different neural network structures was not the main focus of this work. When this work was still in the conceptual stage, much time was invested in evaluating the different classes of networks and their characteristics. By the time the prototype of this system was implemented, the basic architecture of the network had already been decided upon. An iterative approach for the design of the system was taken. When performance was not acceptable, the bottleneck in the system was examined and appropriately modified.

As with any connectionist approach, there are numerous questions that are raised with this work. There are many alternative ways of posing the problem in a connectionist framework. Although the decisions that have been made for this project yield good results, I do not claim that they are optimal. While the exact values specified for this system may not be crucial and perhaps other combinations may be equally effective, the system described in this manuscript has proven to give excellent results.

Some design tools were developed to examine and visualize activations and the weights in the network. Initially, this was done to verify that correct operation of the backpropagation simulator. Later, these tools were briefly used to try to gain insight into how the network was solving the classification task. In fact, this is one of the potential advantages of a connectionist approach. Unfortunately, to date there are no

general purpose methods for examining the internal representations that are formed in the network. The most popular approach, is what I refer to as the hypothesize and test approach. The basic strategy is that certain hypothesis are developed and then the representations are examined to see if the hypothesis can be verified.

Initial efforts to examine the internal structure of the networks developed for the S/G/B/M classification proved evasive. Some simple items were examined first. For example, if most of the weights connecting a particular information bearing parameter to the network are relatively close to zero, then this parameter is not contributing much to the network. Unfortunately, this type of analysis contributed little useful information. It was determined based on the experience of others [117] that any further analysis of this type for this study would not be fruitful and was beyond the scope of this project.

6.6 EFFECTS OF POSTPROCESSING

All experiments were run both with and without the postprocessing stage. This careful examination was useful in isolating problems with the postprocessing stage. Unfortunately, in the current implementation postprocessing only improves performance by roughly 1 percent. The major advantage of the postprocessing stage, is that a "smoother" stream of classifications are obtained. This is best illustrated with a specific example. In figure 6-2 five different traces are shown for the subject JH reciting the utterance "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow." The bottom three traces are from the manual classi-

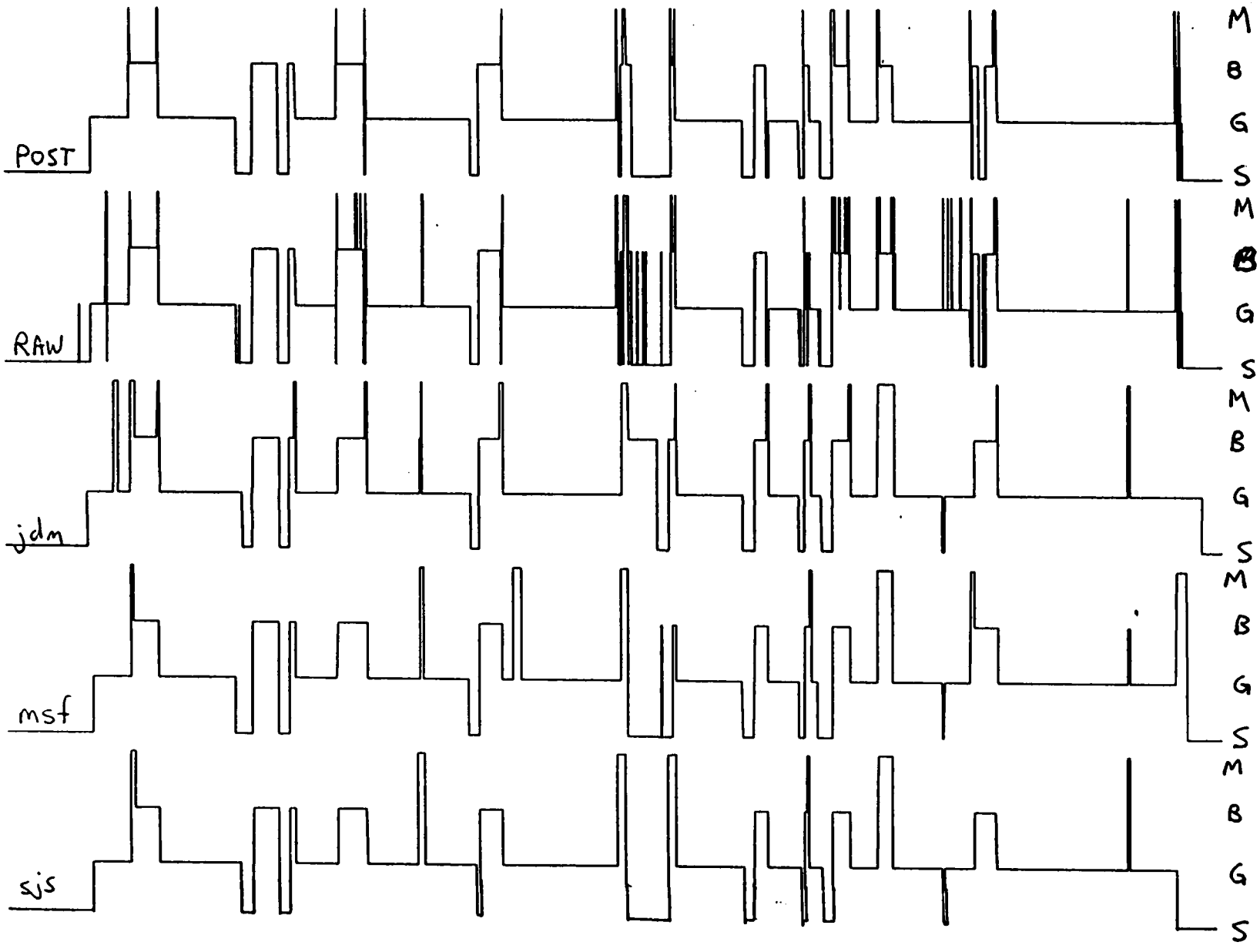


Figure 6-2

fications from the three scientists, the fourth one is for the neural network without postprocessing (RAW) and the top one is for the neural network with postprocessing (POST). The length of each trace is 5000 classifications (5 seconds). One of the most significant effects of postprocessing is that the trace is visually more pleasing than the trace without any postprocessing. For reference all five traces agree unanimously on 83.6 percent of the classifications. Also the percent agreement between SJS and the RAW trace are 91.1 percent and 95.4 percent for the SIMPLE and HAMMING methods respectively, while the SJS and POST traces have SIMPLE and HAMMING agreements of 91.8 and 95.9 percent. From a simple visual inspection, the increment in performance for using postprocessing is typically higher than the increment actually obtained.

Another method that is useful in examining the effects of postprocessing is to count the number of segments that each scientist or classifier obtained. Typically, without postprocessing the classifier produced many more segments than existed in the manually labeled signal. This trend is shown by using the same utterance that was used in the previous example. The data in table 6.26 show that postprocessing dramatically reduces the number of segments obtained from the neural network. After postprocessing the number of segments is in the same range as the manually labeled signals.

6.7 UNRESOLVED ISSUES AND FUTURE DIRECTIONS

In this section, various unresolved issues are discussed along with future directions.

Table 6.26: Effects of postprocessing on the number of segments for a single utterance spoken by subject JH

Expert	Class			
	S	G	B	M
SJS	10	14	9	7
MSF	11	15	12	8
JDM	11	15	12	8
RAW	54	38	65	51
POST	16	14	17	17

Clearly, the segments that the system has the most problem classifying are the mixed segments. If the performance of this system is to be substantially improved, the effort should focused on the mixed category. This could be approached from two different directions. First, alternative information bearing parameters might be developed that are better at classifying mixed segments. Second, it is possible that supplying the neural network with information bearing parameters from multiple frames would assist the network in the classification procedure.

Although this system has shown impressive results, this study has been limited to two native midwestern speakers. It was demonstrated that the networks that were cross validated with different speakers did not suffer major performance degradation. Unfortunately, two speakers is probably too few to claim speaker independence. The design of the system has remained as general purpose as possible and an concerted effort was made not to take advantage of speaker dependent information. If it turns out that the networks developed in this study are speaker dependent, it is my belief that the problem with multiple speakers could be simply addressed by retraining the network.

This study was performed with excellent recording conditions. It would be interesting to evaluate the performance of this system with different recording conditions. The assumption is that the system is quite sensitive to environmental conditions. The hope is that modifying the recording conditions (i.e., noisy environment, telephone quality speech) would simply require retraining the network to achieve results similar to the maximum agreements between speech scientists in the same conditions.

7. SUMMARY

A system has been developed in an attempt to solve a basic problem in speech processing. This work is the first reported attempt to automatically classify an acoustic signal into four categories based on the location of the sound sources in the vocal tract. The four categories are referred to as silence (S), glottal source (G), burst friction (B), and mixed (M). Although similar subproblems have been addressed, it appears that these previous attempts all involve substantial simplifications to the problem that is being investigated in this study. The difficulty of this problem is well summarized by a quotation about a subset of the problem I am concerned with from the head of the Speech Research Department at AT&T Bell Laboratories and another well known speech scientist, "The problem of reliably discriminating among voiced speech, unvoiced speech, and silence is one of the most difficult problems in speech analysis" [37].

A four stage model has been built and extensively tested for the S/G/B/M classification. In the first stage, the speech signal is preprocessed to transform the signal into a form suitable for the extraction of information bearing parameters. Second, a set of 14 information bearing parameters are extracted from the signal. Next, these parameters are used as input to a feed-forward neural network that classifies the signal into one of four categories. Finally, the stream of classifications is filtered through a postprocessor to correct obvious mistakes.

Using this framework, it was shown that a connectionist system could be taught to correctly classify speech into four categories (S/G/B/M). Many experiments were run in an effort to design the most successful classifier. The best system was one that used a three layer feed-forward network. This network was trained using the backpropagation learning algorithm on the first 34 seconds of the rainbow passage from a female speaker. The accuracy on the training set was 90.8 percent. When tested on the same speaker reciting the remaining 87 seconds of the rainbow passage the accuracy was 90.0 percent. The same network was also tested with a male speaker reciting the full rainbow passage. The accuracy on this 106 second passage was 88.8 percent. In light of the expected percent agreement between speech scientists (91.5), these results are extremely promising.

Using a multi-disciplinary approach to address a fundamental problem in speech processing, this work not only has immediate and direct applications to speech recognition, but it will also be useful for both speech analysis and synthesis.

8. ACKNOWLEDGMENTS

My sincere appreciation goes to my both thesis advisors. Dr. James D. Miller was the catalyst that sparked my interest in speech. His enthusiasm and positive attitude were always balanced with patience and encouragement. Under his supervision I have been able to gain invaluable practical experience, and I have also benefited from the latitude he has given me to explore many seemingly unrelated topics that have strengthed my background. Dr. Jerome R. Cox has not only provided advice, but he has been responsible for building a computer science department that not only deserves international respect, but that has also earned my respect. His influence is reflected in many aspects of my training.

For the countless conversations, suggestions, and thought provoking comments, I must thank my thesis committee; Dr. Maynard Engebretson, Dr. Stan Kwasny, and Dr. Takayuki Dan Kimura. Additionally, I would like to thank to the members of the Speech Group and the many other scientists at Central Institute for the Deaf who have not only helped me with work for my degree, but have also forced me to think about science in general.

I also owe thanks to both my parents for all their love and support throughout my lengthy academic career.

Finally, I would like to thank my wife, Alyssa. Without her, this degree, along with the rest of my life would not be what they are today.

APPENDICES

APPENDIX A

The Rainbow Passage.

When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

Throughout the centuries, men have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Other men have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflections of the sun's rays by the rain. Since then physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbow. Many complicated

ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the water drops, and the width of the colored band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of superposition of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green lights when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.

BIBLIOGRAPHY

[1] J. L. Flanagan and L. Cherry. Excitation of vocal-tract synthesizers. *Journal of the Acoustical Society of America*, 45(3):764–769, 1968.

[2] Gunnar Fant. *Acoustic Theory of Speech Production*. Mouton and Co., Hague, Netherlands, 1960.

[3] Gloria J. Borden and Katherine S. Harris. *Speech Science Primer*. Williams and Wilkins, Baltimore, MD, 1984.

[4] James L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, New York, NY, 1972.

[5] Richard P. Lippmann. Neural network classifiers for speech recognition. *Lincoln Laboratory Journal*, 1(1):107–124, 1988.

[6] D. R. Reddy. Segmentation of speech sounds. *Journal of the Acoustical Society of America*, 40(2):307–312, 1966.

[7] Clifford J. Weinstein, Stephanie S. McCandless, Lee F. Mondshein, and Victor W. Zue. A system for acoustic-phonetic analysis of continuous speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):54–67, February 1975.

[8] Richard Schwartz and John Makhoul. Where the phonemes are: Dealing with ambiguity in acoustic-phonetic recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):50–53, February 1975.

[9] K. K. Paliwal and P. V. S. Rao. Acoustic phonetic recognition of continuous speech. *9th International Congress of Acoustics, Spain*, 1977.

[10] P. Regel. A module for acoustic-phonetic transcription of fluently spoken German speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-30:440–450, 1982.

[11] K. K. Paliwal and P. V. S. Rao. Synthesis-based recognition of continuous speech. *Journal of the Acoustical Society of America*, 71:1016-1021, 1982.

[12] Hubert W. Upton. Wearable eyeglass speechreading aid. *Gallaudet Conference on Speech Aids - American Annals of the Deaf*, 113(2):222–229, March 1968.

[13] Dennis H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3):971–995, March 1980.

[14] John Makhoul, R. Viswanathan, Richard Schwartz, and A. W. F. Huggins. A mixed-source model for speech compression and synthesis. *Journal of the Acoustical Society of America*, 64(6):1577–1581, December 1978.

[15] Ronald W. Schafer and Lawrence R. Rabiner. Design and simulation of a speech analysis-synthesis based on short-time Fourier analysis. *IEEE Transactions on Audio and Electroacoustics*, AU-21(3):165–174, June 1973.

[16] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House. Reduction of speech spectra by analysis-by-synthesis techniques. *Journal of the Acoustical Society of America*, 33(12):1725–1736, December 1961.

[17] B. Atal and J. Remde. A new model of LPC excitation for producing natural-sounding speech at low bit rate. *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing*, pages 614–617, 1982.

[18] M. H. Savoji. A robust algorithm for accurate endpointing of speech signals. *Speech Communication*, 8(1):45–60, March 1989.

[19] LaDeanna F. Weigelt, Steven J. Sadoff, and James D. Miller. An algorithm for distinguishing between voiceless stops and voiceless fricatives. *Journal of the Acoustical Society of America*, 85 Supplement 1, May 1989.

[20] LaDeanna F. Weigelt, Steven J. Sadoff, and James D. Miller. An algorithm for distinguishing between voiced stops and voiced fricatives. *Journal of the Acoustical Society of America*, 86 Supplement 1, November 1989.

[21] LaDeanna F. Weigelt, Steven J. Sadoff, and James D. Miller. Plosive/fricative distinction: The voiceless case. *Journal of the Acoustical Society of America*. In Press.

[22] James D. Miller. Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America* [illegible]):2114–2134, May 1989.

[23] James D. Miller. Auditory-perceptual processing of speech waveforms. In William A. Yost and Charles S. Watson, editors, *Auditory Processing of Complex Sounds*, pages 257–266. Lawrence Erlbaum Associates, Hillsdale, N.J, 1987.

[24] Pierre C. Delattre, Alvin M. Liberman, and Franklin S. Cooper. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27(4):769–773, 1955.

[25] Alvin Liberman, Katherine Safford Harris, Peter Eimas, Leigh Lisker, and Jarvis Bastian. An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language and Speech*, 4:175–195, 1961.

[26] Michael F. Dorman, Lawrence J. Raphael, and Alvin M. Liberman. Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, 65(6):1518–1532, June 1979.

[27] Lori F. Lamel, Lawrence R. Rabiner, Aaron E. Rosenberg, and Jay G. Wilpon. An improved endpoint detector for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(4):777–785, August 1981.

[28] E. F. O'Neill. TASI. *Bell Labs Record*, 37:83–87, March 1959.

[29] H. Miedema and M. G. Schachtman. TASI quality – effect of speech detectors and interpolation. *The Bell System Technical Journal*, 51:1455–1473, July 1962.

[30] E. Fariello. A novel digital speech detector for improving effective satellite capacity. *IEEE Transactions on Communications*, COM-20:55 60, February 1972.

[31] P. G. Drago, A. M. Molinari, and F. C. Vagliani. Digital dynamic speech detectors. *IEEE Transactions on Communications*, COM-26(1):140-145, January 1978.

[32] Bernard Gold. Note on buzz-hiss detection. *Journal of the Acoustical Society of America*, 36(9):1659–1661, September 1964.

[33] A. Michael Noll. Cepstrum pitch detection. *Journal of the Acoustical Society of America*, 41(2):293–309, February 1967.

[34] Osamu Fujimura. An approximation to voice aperiodicity. *IEEE Transactions on Audio and Electroacoustics*, AU-16(1):68–72, March 1968.

[35] Vivien C. Tartter. What's in a whisper? *Journal of the Acoustical Society of America*, 86(5):1678–1683, November 1989.

[36] B. S. Atal and L. R. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(3):201–212, June 1976.

[37] Lawrence R. Rabiner and Marvin R. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-25(4):338–343, August 1977.

[38] Chong Kwan Un and Hyeong Ho Lee. Voiced/unvoiced/silence discrimination of speech by delta modulation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28(4):398–407, August 1980.

[39] Leah J. Siegel and Alan C. Bessey. Voiced/unvoiced/mixed excitation classification of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-30(3):451–460, June 1982.

[40] Siegfried G. Knorr. Reliable voiced/unvoiced decision. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(3):263–267, June 1979.

[41] Hideki Kasuya and Hisashi Wakita. An approach to segmenting speech into vowel- and nonvowel-like intervals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(4):319–327, August 1979.

[42] Hisao M. Chang. *SWIS: See What I Say A Speaker-Independent Word Recogntion System by Phoneme-Oriented Mapping on a Phonetically Encoded Auditory-Perceptual Speech Map.* PhD thesis, Washington University, Department of Computer Science, 1987.

[43] Alex Waibel. *Prosody and Speech Recognition.* Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.

[44] Grant Fairbanks. *Voice and Articulation Drillbook.* Harper & Brothers, Publishers, New York, NY, second edition, 1960.

[45] J. D. Markel and A. H. Gray, Jr. *Linear Prediction of Speech.* Springer-Verlag, New York, NY, 1982.

[46] B. A. Dautrich, L. R. Rabiner, and T.B. Martin. The effects of selected signal processing techniques on the performance of a filter-bank-based isolated word recognizer. *The Bell System Technical Journal*, 62(5):1311–1336, May 1983.

[47] K. K. Paliwal. Effect of preemphasis on vowel recognition performance. *Speech Communication*, 3(1):101–106, April 1984.

[48] Ralph K. Potter, George A. Kopp, and Harriet C. Green. *Visible Speech.* D. Van Nostrand, New York, NY, 1947.

[49] A. J. Presti. High-speed sound spectrograph. *Journal of the Acoustical Society of America*, 40:628–634, 1966.

[50] Michael D. Riley. *Speech Time-Frequency Representations.* Kluwer Academic Publishers, Norwell, MA, 1989.

[51] Fredric J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, January 1978.

[52] J. F. Kaiser. Nonrecursive digital filter design using the i_0-sinh window function. *Proceedings of the 1974 IEEE Int. Symp. on Circuits and Syst.*, pages 20–23, April 1974.

[53] Gordon E. Peterson and Harold L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184, March 1952.

[54] C. J. Darwin and Mark Pearson. What tells us when voicing has started. *Speech Communication*, 1(1):29–44, May 1982.

[55] L. R. Rabiner and M. R. Sambur. An algorithm for determining the endpoints of isolated utterances. *Bell System Technical Journal*, 54(2):297–315, February 1975.

[56] Harvey Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, 5:82–108, October 1933.

[57] S. S. Stevens. Perceived level of noise by Mark VII and decibels (E). *Journal of the Acoustical Society of America*, 51(2):575–601, 1972.

[58] K. K. Paliwal, S. S. Sinha, and A. Agarwal. An isolated word recognition system for Hindi digits using linear time normalisation. *Journal of the Institution of Electronics and Telecom. Engineers*, 29(1):18–22, January 1983.

[59] J. C. R. Licklider and Irwin Pollack. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *Journal of the Acoustical Society of America*, 20(1):42–51, January 1948.

[60] J. C. R. Licklider. The intelligibility of amplitude-dichotomized, time-quantized speech waves. *Journal of the Acoustical Society of America*, 22(6):820–823, November 1950.

[61] R. J. Niederjohn, M. W. Krutz, and B. M. Brown. An experimental investigation of the perceptual effects of altering the zero-crossings of a speech signal. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(5):618–625, May 1987.

[62] Russell J. Niederjohn. A mathematical formulation and comparison of zero-crossing analysis techniques which have been applied to automatic speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(4):373–379, August 1975.

[63] Russell J. Niederjohn and Meir Lahat. A zero-crossing consistency method for formant tracking of voiced speech in high noise levels. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(2):349–355, April 1985.

[64] B. P. Bogert, M. J. R. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes. In M. Rosenblatt, editor, *Proceedings of the Symposium on Time Series Analysis*, pages 209–243, New York, NY, 1963. John Wiley & Sons.

[65] B. S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):637–655, August 1971.

[66] Tirupattur V. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(4):309–319, August 1979.

[67] Arun Kumar, Danial R. Fuhrmann, Michael Frazier, and Björn Jawerth. A new transform for time-frequency analysis. Technical Report WUCS-89-27, Washington University, Computer Science Department, October 1989.

[68] Arun Kumar and Danial R. Fuhrmann. The Frazier-Jawerth transform. In *ICASSP 90 Proceedings*. IEEE, April 1990.

[69] Andrew Wilson Howitt. Application of the Wigner distribution to speech analysis. Based on his Master's Thesis, 1987.

[70] E. Zwicker, E. Terhardt, and E. Paulus. Automatic speech recognition using psychoacoustic models. *Journal of the Acoustical Society of America*, 65(2):487–498, 1979.

[71] Dennis H. Klatt. Speech processing strategies based on auditory models. In Rolf Carlson and Bjorn Granstrom, editors, *The Representation of Speech in the Peripheral Auditory System*, pages 181–196, Amsterdam, 1982. Elsevier Biomedical Press.

[72] Rolf Carlson and Bjorn Granstrom. Towards an auditory spectograph. In Rolf Carlson and Bjorn Granstrom, editors, *The Representation of Speech in the Peripheral Auditory System*, pages 109–114, Amsterdam, 1982. Elsevier Biomedical Press.

[73] R. L. Wegel and C. E. Lane. The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Physics Review*, 23:266–285, 1924.

[74] Lloyd A. Jeffress. Masking. In Jerry V. Tobias, editor, *Foundations of Modern Auditory Theory*, pages 85–114. Academic Press, New York, 1970.

[75] Harvey Fletcher and W. A. Munson. Relation between loudness and masking. *Journal of the Acoustical Society of America*, 9(1):1–10, July 1937.

[76] E. Zwicker, G. Flottrop, and S. S. Stevens. Critical band width in loudness summation. *Journal of the Acoustical Society of America*, 29(5):548–557, May 1957.

[77] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *Journal of the Acoustical Society of America*, 33(2):248, February 1961.

[78] Brian C. J. Moore and Brian R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3):750–753, September 1983.

[79] Betram Scharf. Critical bands. In Jerry V. Tobias, editor, *Foundations of Modern Auditory Theory*, pages 157–202. Academic Press, New York, 1970.

[80] Tammo Houtgast. *Lateral Suppresion in Hearing.* PhD thesis, Vrije University, Amsterdam, 1974.

[81] Robert H. Gilkey and Donald E. Robinson. Models of auditory masking: A molecular psychophysical approach. *Journal of the Acoustical Society of America*, 79(5):1499–1510, May 1986.

[82] R. H. Gilkey. Spectral and temporal comparisons in auditory masking. In William A. Yost and Charles S. Watson, editors, *Auditory Processing of Complex Sounds*, pages 26–36. Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[83] David M. Green, editor. *Profile Analysis.* Oxford University Press, New York, NY, 1988.

[84] Julius L. Goldstein. Updating cochlear driven models of auditory perception: A new model for nonlinear auditory frequency analysing filters. In *Working Models of Human Perception*, pages 19–57. Academic Press, London, 1988.

[85] B. A. Dautrich, L. R. Rabiner, and T.B. Martin. On the effects of varying filter bank parameters on isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-31(4):793-806, August 1983.

[86] George M. White and Richard B. Neely. Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(2):183–188, April 1976.

[87] Laveen N. Kanal. Interactive pattern analysis and classification systems: A survey and commentary. *Proceedings of the IEEE*, 60:1200-1215, October 1972.

[88] Richard P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4(2):4–22, April 1987.

[89] Igor Aleksander, editor. *Neural Computing Architectures: the Design of Brain-Like Machines.* MIT Press, Cambridge, MA, 1989.

[90] Scott E. Fahlman and Geoffrey E. Hinton. Connectionist architectures for artificial intelligence. *IEEE Computer*, 21(3):100–108, January 1987.

[91] David E. Rumelhart, James L. McClelland, and The PDP Research Group. *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, volume 1: Foundations. MIT, Cambridge, MA, 1987.

[92] R. Colin Johnson and Chappell Brown. *Cognizers: Neural Networks and Machines That Think*. John Wiley & Sons, New York, NY, 1988.

[93] Terrence J. Sejnowski and Charles R. Rosenberg. NETtalk: a parallel network that learns to read aloud. Technical Report JHU/EECS-86/01, Johns Hopkins University, 1986.

[94] Patricia S. Churchland and Terrence J. Sejnowski. Perspectives on cognitive neuroscience. *Science*, 242:741–745, November 1988.

[95] Jerome A. Feldman, Mark A. Fanty, Nigel H. Goddard, and Kenton J Lynne. Computing with structured connectionist networks. *Communications of the ACM*, 31(2):170–187, February 1988.

[96] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, NY, 1973.

[97] Dennis H. Klatt. Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America*, 82(3):737–793, September 1987.

[98] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In David E. Rumelhart, James L. McClelland, and The PDP Research Group, editors, *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, volume 1: Foundations. MIT, Cambridge, MA, 1987.

[99] Teuvo Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, New York, NY, 1984.

[100] Teuvo Kohonen. The "neural" phonetic typewriter. *IEEE Computer*, 21(3):11–22, March 1988.

[101] Candace A. Kamm, Lynn A. Streeter, Yana Kane-Esrig, and David J. Burr. Comparing performance of spectral distance measures and neural network methods for vowel recognition. *Computer Speech and Language*, 3:21–34, 1989.

[102] Alex Waibel, Hidefumi Sawai, and Kiyohiro Shikano. Modularity and scaling in large phonemic neural networks. Technical Report TR-I-0034, ATR Interpreting Telephony Research Laboratories, August 1988.

[103] Jeffrey L. Elman and David Zipser. Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, 83(4):1615–1626, April 1988.

[104] D. E. Rumelhart, G. E. Hinton, and J. L. McClelland. A general framework for parallel distributed processing. In David E. Rumelhart, James L. McClelland, and The PDP Research Group, editors, *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, volume 1: Foundations. MIT, Cambridge, MA, 1987.

[105] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. Technical Report ICS-8805, University of California, San Diego, Institute for Cognitive Science, October 1988.

[106] Marvin Minsky and Seymour Papert. *Perceptrons*. MIT Press, Cambridge, MA, 1969.

[107] Takayuki Dan Kimura. A learning algorithm for acyclic neural networks. Technical Report WUCS-89-26, Washington University, Computer Science Department, 1989.

[108] R. Paul Gorman and Terrence J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1, 1988.

[109] T. D. Kimura, 1989. Personal Communication.

[110] Josiah C. Hoskins. Speeding up artificial neural networks in the "real" world. Technical Report STP-049-89, Microelectronics and Computer Technology Corporation, January 1989.

[111] L. R. Rabiner, Marvin R. Sambur, and Carolyn E. Schmidt. Applications of a nonlinear smoothing algorithm to speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(6):552–557, December 1975.

[112] Frederick P. Brooks, Jr. *The Mythical Man-Month*. Addison-Wesley, Reading, MA, 1982.

[113] R. H. Gilkey and M. E. Partridge. Direct memory-access control of the Micro Technology Unlimited DigiSound-16 with a Q-bus based computer. In preparation, 1990.

[114] Carver Mead. *Analog VLSI and Neural Systems*. Addison Wesley, Reading, MA, 1989.

[115] N. H. Goddard, M. A. Fanty, and K. Lynne. The Rochester connectionist simulator. Technical Report 233, University of Rochester, Computer Science Department, 1987.

[116] K. J. Lang and M. J. Witbrock. Learning to tell two spirals apart. In D. S. Touretzky, G. E. Hinton, and T. J. Sejnowski, editors, *Proceedings of the 1988 Connectionist Models Summer School*, San Mateo, CA, 1988. Morgan Kaufmann Publishers.

[117] T. D. Kimura, 1990. Personal Communication.

11. VITA

Biographical items on the author of the thesis, Mr. S. J. Sadoff

1. Born October 23, 1963.
2. Attended Washington University from August, 1981 to May, 1985. Received the degree of Bachelor of Science in Computer Science in May, 1985.
3. Attended Washington University from August, 1985 to May, 1987. Received the degree of Masters of Science in Electrical Engineering in May, 1985.
4. Attended Washington University from August, 1987 to the present date.
5. Membership in Professional and Honor Societies: Tau Beta Pi, I.E.E.E., A.C.M., A.S.A., and NSSLHA.

May, 1990

Washington
WASHINGTON·UNIVERSITY·IN·ST·LOUIS

Date May 16, 1990

To the Graduate School:

The dissertation and dissertation abstract of John Warren Hawks

entitled Perceptual Aspects of a Three-dimensional Vowel Space

have been examined by the undersigned and have our approval for acceptance in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

We further recommend bringing the dissertation to oral defense.

Signed _____
Chairman, Dissertation Advisory Committee

_____
Member, Dissertation Advisory Committee

_____
Member, Dissertation Advisory Committee

_____

I concur in this recommendation.

_____
Director of Communication Sciences

_____
Chairman of Department

This form should accompany the dissertation when it is submitted to the Graduate School of Arts and Sciences, 211A South Brookings Hall.

AFOSR Grant G-AFOSR-86-0335
Final Technical Report
Appendix

C

Graduate School of Arts and Sciences

Washington University

St. Louis, Missouri 63130

Date May 16, 1990

TO THE GRADUATE COUNCIL:

We, the undersigned, report that as a committee we have examined JOHN WARREN HAWKS upon the work done in the subjects named below:

Major COMMUNICATION SCIENCES

and find that (his, ~~her~~) attainments (are, are not) such that (he, ~~she~~) may properly be admitted to the degree of DOCTOR OF PHILOSOPHY (with, ~~without~~) thesis.

Chairman

I dissent from the foregoing report.

One copy is to be prepared, signed, and forwarded to the Office of the Dean.

Date Recorded: Office of the Dean

WASHINGTON UNIVERSITY
Department of Speech and Hearing
Program in Communication Sciences

Dissertation Committee:
James D. Miller, Chairman
Ira J. Hirsh
A. Maynard Engebretson
Margaritis S. Fourakis

Perceptual Aspects of a Three-dimensional Vowel Space

by

John Warren Hawks

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August, 1990
Saint Louis, Missouri

Abstract

Chairman: James D. Miller

One of the methods that has been utilized for vowel classification in natural speech posits a three-dimensional space, defined by certain acoustic characteristics of the vowels related to the first three significant prominences, or formants, in their short-term spectra ($F1$, $F2$, and $F3$) and voice pitch ($F0$). Within this space, productions of like vowels are mapped onto subspaces, called target zones. Here, the characteristics of this space and its subspaces were studied by means of subject responses to synthetic vowel-like tokens representing unique points in this space. The first experiment utilized the identifications of eight listeners to construct a vowel map of this space. Vowel categories for American English can be represented as abutting and non-overlapping target zones which correctly classify over 99% of the plurality-based identifications. The first two formants ($F1$ and $F2$) appeared to be the primary determinants in the identification of non-retroflex vowels. The third formant ($F3$) determined the perception of retroflex (r-colored) vowels, and contributed to the phonetic saliency of other vowel categories. Furthermore, phonetic saliency varied in an orderly way with location within a vowel target zone.

A second experiment investigated the discrimination of complex sounds represented as points in the three-dimensional vowel space by estimating difference limens (DLs), expressed as distance in the space, for synthetic, vowel-like tokens. Four subjects were used to estimate DLs along 102 straight-line continua. Continua emanated in six directions from 17 locations in the space. Movement along continua resulted in distinct, multi-formant patterns of frequency change which varied with the direction and axis of movement. The average DL for distance across all continua was estimated to be .01 log units, but DLs were significantly different for the three axes of the space. Considerable variation found in DLs associated

with individual reference points may be related to differences in reference formant patterns. The DL results for multiple-formant-change continua were found to be significantly smaller than similar DL estimates for single-formant-change continua. Many of these differences could be accounted for by an additive model whereby each changing formant contributes independent information to perception.

Acknowledgements

Many people have played many roles in aiding and abetting this dissertation. To anyone of these people whom I unintentionally omit from these acknowledgements, I thank you. I would first and foremost like to thank my wife for her ever-enduring tolerance of this project, the many extra hours it took me away from her, the completion postponements, and my temporary insanity. I only hope I can repay this debt. I also thank my children, Ben, Adam, and the little one in the "oven" for constantly reminding me of what life is really about. I hope that now I can start living it with you more fully.

I would like to thank and acknowledge my dissertation committee, Drs. James D. Miller, Ira Hirsh, Maynard Engebretson, and Marios Fourakis, for their help, advice, and encouragement. Dr. Miller, my advisor, deserves special thanks in that he has been instrumental in shaping my education and career since first meeting with me in 1982 to discuss graduate school. He has led me, advised me, employed me, and taught me much about the right and the wrong ways to do things in this business. Special thanks also go to my best friend and officemate, Marios Fourakis for all the hours spent sharing what he knows and keeping me on track.

Although I am pleased that most of the computer programming required of this project I was able to do for myself, thanks go to Steven J. Sadoff and Frank Kramer for their invaluable assistance and to the late Dennis Klatt for providing the world with a wonderful tool in his software synthesizer. I am extremely grateful for the assistance and discussion on statistical issues from Caroline B. Monahan, Martha Storandt, and Janet Weisenberger, and on psychophysical issues from Bob Gilkey. Thanks also go to the CID technical support staff, headed by Arnold Heidbreder, for their ongoing efforts to ensure that no project is short-circuited.

A special acknowledgement to my fellow doctoral students, Chip Nicholas, Punita Singh, Steve Sadoff, and Lyn Shields, for their comradery during this whole ordeal. Thanks also to all those who served as subjects for the pilot work and the experiments for their countless hours of listening, as well as to another officemate, Dr. Michael Gottfried, for numerous discussions and encouragement.

This work was supported by a grant from NIDCD.

Contents

1 Introduction and background 18

1.1 Introduction 18

1.2 The Auditory-Perceptual Theory (*APT*) 20

1.2.1 Auditory-perceptual space (*APS*) 20

1.2.2 Formant location in the *APS* 21

1.2.3 Concept and estimation of perceptual target zones (*PTZs*) 23

1.3 Overview of experiments 31

2 *Experiment I:* Perceptual Mapping of the APS Vowel Space. 32

2.1 Introduction 32

2.2 Methods 36

2.2.1 Stimuli 36

2.2.2 Procedure 43

2.2.3 Subjects 44

2.3 Results 45

2.3.1 General observations 45

2.3.2 Identifications 46

2.3.3 Ratings 48

2.3.4 Synthetic speech-based (*SSB*) target zones 55

2.3.5 Qualitative analysis of synthetic speech-based target zones 63

2.3.6 Plurality agreements on identifications 66

2.3.7 Confidence ratings 67

2.3.8 Individual differences in identification responses 75
2.3.9 Linear Discriminant Analyses 92
2.3.10 Agreement by z' plane 95
2.4 Comparisons of vowel classification schemes 99
2.4.1 $F1$ x $F2$ 100
2.4.2 Comparison of synthetic and natural speech-based target zones . . . 110
2.4.3 Classification using bark differences 120
2.4.4 Vowel classification utilizing extrinsic specification 122
2.5 Summarization and Discussion of Experiment I 126

3 Experiment II: Estimation of difference limen for distance (d) in the APS vowel space 134

3.1 Introduction 134
3.2 Methods 136
3.2.1 Stimuli 136
3.2.2 Procedure 144
3.2.3 Apparatus 145
3.2.4 Subjects 145
3.2.5 Training 145
3.2.6 Formant change and *APS* continua 146
3.3 Results 149
3.3.1 General Analyses 149
3.3.2 Analyses by reference group 151
3.3.3 Analyses by Subject 153
3.3.4 Analyses by Percentage of Formant Change 154
3.3.5 Analyses of Movement in z' 157
3.3.6 Overall Discrimination by Reference 158
3.3.7 Single vs. Multiple Formant Movement 161
3.4 Summarization and Discussion of Experiment II 165

4 Final Comments and Implications for Future Research 174

A Formant location in the *APS* **188**

B Synthesis Parameter Specifications **198**

C Spectral Envelopes for Experiment II reference tokens **203**

List of Tables

1.1 ARPAbet symbols for representing the phoneme-like units of English within a computer. (From Lee and Shoup, 1980) 25

2.1 Frequency of ID responses by subject response set. 47
2.2 Percentages of identification agreement by subject response set. 49
2.3 Frequency of rating responses by subject response set. 51
2.4 Percentages of agreement on confidence ratings by subject response set. . . 54
2.5 Agreement on identification responses by plurality frequency. 66
2.6 Linear discriminant analyses of plurality identifications. 93
2.7 Linear discriminant analyses of plurality identifications (no /ER/). 94
2.8 Averaged values of minimum, maximum, and average z' for CID Natural Speech Database and results from Peterson and Barney (1952). 96
2.9 Average percentages of pair-wise agreements for all subject response sets by z' range. 97
2.10 Preliminary vowel classification using *NSB* and *SSB* target zones. 111
2.11 Corrected vowel classification using *NSB* and *SSB* target zones. 112
2.12 Classification using *NSB*, *SSB*, and *HiR SSB* target zones. 118
2.13 Vowel feature system using bark-difference dimensions from Syrdal and Gopal (1986). Features in parentheses are based on best fit to synthetic data. . . . 121
2.14 Vowel classification using Syrdal and Gopal (1986) classification scheme. . . 121
2.15 Vowel classification using Neary (1977) classification scheme. 126

3.1 Mean DL in log unit distance for various conditions across subjects and replications. 150

3.2 Probabilities of significance for factors from overall and individual reference group analyses-of-variance of DLs expressed as distance. 151
3.3 Ranked DL results associated with each center reference point. 152
3.4 Ranked DL results associated with each ambiguous reference point. 153
3.5 Analysis of variance results for effect of conditions overall and by subject. . 154
3.6 Significances of factors from overall and individual reference group analyses-of-variance of DLs expressed as percent F2 change. 156
3.7 Analysis-of-variance results for effect of conditions overall and by reference group for z' continua. 158
3.8 R^2 values from multiple regression analyses of DL results and specified variable sets (See text). 172

B.1 Synthesis parameter specifications. 199
B.2 Time-varying synthesis parameter specifications for F0 (x10) and amplitude. 200
B.3 Formant bandwidths (BW) by formant frequency (Frmt) in Hertz utilized for all synthetic tokens in all experiments. 201

C.1 Formant ($F1$, $F2$, $F3$) values for the 17 reference points used in Experiment II. 204

List of Figures

1-1 (a) View of vowel "slab" (gridded area) in *APS* dimensions; (b) view of the same plane in transformed (x', y', z') dimensions. 22

1-2 Estimations of perceptual target zones for American English in APS $x'y'$ coordinates based on measurements of 435 vowels from natural speech. . . . 26

1-3 Estimations of perceptual target zones for American English in APS $x'y'$ coordinates based on measurements of 2051 vowels from natural speech. . . 27

1-4 Estimations of perceptual target zones for American English in APS $y'z'$ coordinates based on measurements of 435 vowels from natural speech. . . . 28

1-5 Estimations of perceptual target zones for American English in APS $y'z'$ coordinates based on measurements of 2051 vowels from natural speech. . . 29

2-1 Subjects' identifications for synthetic vowels from R.L. Miller (1953) plotted in APS $x'y'$ coordinates along with current target zone estimates from Figure 1-3. 35

2-2 Location in APS $x'y'$ coordinates of synthesizable tokens for one z' plane $(z' = 0.700)$. 37

2-3 Location in APS $y'z'$ coordinates of z' planes utilized for Experiment I. . . 38

2-4 Locations in $x'y'$ of tokens in one z' plane $(z' = 0.700)$ acceptable for synthesis after applying formant-range limiting criteria plotted with the most recent target zone estimations from Figure 1-3. 40

2-5 (a) Overall amplitude contour used for token synthesis; (b) Fundamental frequency ($F0$) contour used for token synthesis; (c) Q as a function of the ratio of formant center frequency (Fc) over fundamental frequency ($F0$) used for formant bandwidth calculation in token synthesis (See text). 42
2-6 Mean agreement between each subject response set and all other response sets on identification of the 1725 synthetic vowels. Error bars indicate ± 1 standard deviation. 50
2-7 Percentage of subjects' confidence rating responses by individual identification category. 52
2-8 (a) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.80$ plane. 56
2-8 (b) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.75$ plane. 57
2-8 (c) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.70$ plane. 58
2-8 (d) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.65$ plane. 59
2-8 (e) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.60$ plane. 60
2-8 (f) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.55$ plane. 61
2-8 (g) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.50$ plane. 62

2-9 Synthetic-speech-based target zones (solid lines) from Figure 2-8 and natural-speech-based target zones (dashed lines) from Figure 1-3 for the $z' = .70$ plane. 65
2-10 (a) Plurality frequencies (See text) for all tokens in the $z' = 0.80$ plane. . . 68
2-10 (b) Plurality frequencies (See text) for all tokens in the $z' = 0.75$ plane. . . 69
2-10 (c) Plurality frequencies (See text) for all tokens in the $z' = 0.70$ plane. . . 70
2-10 (d) Plurality frequencies (See text) for all tokens in the $z' = 0.65$ plane. . . 71
2-10 (e) Plurality frequencies (See text) for all tokens in the $z' = 0.60$ plane. . . 72
2-10 (f) Plurality frequencies (See text) for all tokens in the $z' = 0.55$ plane. . . . 73
2-10 (g) Plurality frequencies (See text) for all tokens in the $z' = 0.50$ plane. . . 74
2-11 (a) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.80$ plane. 76
2-11 (b) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.75$ plane. 77
2-11 (c) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.70$ plane. 78
2-11 (d) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.65$ plane. 79
2-11 (e) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.60$ plane. 80
2-11 (f) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.55$ plane. 81
2-11 (g) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.50$ plane. 82
2-12 Mean sums of confidence ratings for tokens grouped by plurality frequency. Error bars indicate ± 1 standard deviation. 83
2-13 (a) Locations of tokens for which identifications agreed across three response sets for subject 1F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications. . . . 85

2-13 (b) Locations of tokens for which identifications agreed across three response sets for subject 2F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications. . . . 86
2-13 (c) Locations of tokens for which identifications agreed across three response sets for subject 3F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications. . . . 87
2-13 (d) Locations of tokens for which identifications agreed across three response sets for subject 5F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications. . . . 88
2-13 (e) Locations of tokens for which identifications agreed across three response sets for subject 3M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications. . . . 89
2-13 (f) Locations of tokens for which identifications agreed across three response sets for subject 4M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications. . . . 90
2-13 (g) Locations of tokens for which identifications agreed across three response sets for subject 5M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications. . . . 91
2-14 Vowel ellipses plotted in $F1$ x $F2$ space from Figure 8 of Peterson and Barney (1952). . . . 101
2-15 All plurality identifications with ellipses from Figure 2-14 in $F1$ x $F2$ space. 102
2-16 (a) Plurality identifications from the $z' = 0.80$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space. . . . 103
2-16 (b) Plurality identifications from the $z' = 0.75$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space. . . . 104
2-16 (c) Plurality identifications from the $z' = 0.70$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space. . . . 105
2-16 (d) Plurality identifications from the $z' = 0.65$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space. . . . 106

2-16 (e) Plurality identifications from the $z' = 0.60$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space. 107
2-16 (f) Plurality identifications from the $z' = 0.55$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space. 108
2-16 (g) Plurality identifications from the $z' = 0.50$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space. 109
2-17 (a) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.75$ plane which were misclassified by the *SSB* target zones. 115
2-17 (b) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.70$ plane which were misclassified by the *SSB* target zones. 116
2-17 (c) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.65$ plane which were misclassified by the *SSB* target zones. 117
2-18 Location in APS $x'y'$ coordinates of computer-constructed high-resolution target zones based on data from Miller and Hawks, 1989. 119
2-19 Locations in APS $x'y'$ coordinates of EXEMPLARY (circles) and AVERAGE (triangles) vowel reference frameworks used for extrinsic specification in Neary-type classification procedure along with locations (for comparison) of average male vowels from Peterson and Barney, 1952. 124

3-1 Orientation of the six continua in APS $x'y'z'$ coordinates associated with each reference point used in Experiment II. 137
3-2 Locations in APS $x'y'$ coordinates of center references (x's) utilized in Experiment II compared to locations of exemplary reference framework tokens (+'s, See Section 2.4.2) and male average vowels ($\star$'s) from Peterson and Barney, 1952. 140
3-3 Locations in APS $x'y'$ coordinates of points for the first evaluation for ambiguous reference points along with *SSB* target zones for the $z' = 0.70$ plane. 141

3-4 Locations in APS $x'y'$ coordinates of points for the second evaluation for ambiguous reference points along with *SSB* target zones for the $z' = 0.70$ plane. 142

3-5 Locations in APS $x'y'$ coordinates of the ambiguous reference points used in Experiment II along with the *SSB* target zones for the $z' = 0.70$ plane. . . . 143

3-6 Idealized spectra representing the relative shifts in formant frequency for a fixed distance movement along each of the six directional continua (dashed lines) relative to the reference point (solid lines). (a) Continuum 3; (b) Continuum 9; (c) Continuum 12; (d) Continuum 6; (e) Continuum F; (f) Continuum B. 147

3-7 Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for x' continua. 159

3-8 Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for y' continua. 160

3-9 Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for z' continua. 162

3-10 Locations in APS $x'y'$ coordinates of single-formant-change continua relative to multiple-formant-change continua. 163

4-1 Locations of *SSB* target zones in APS $x'y'$ coordinates with axes modified to reflect approximately equal DL units. 179

A-1 Location of seven continua generated in $x'y'$ space with fixed values of $F2$ and $F3$ with $F1$ allowed to vary. 189

A-2 Location of seven continua generated in $x'y'$ space with fixed values of $F1$ and $F3$ with $F2$ allowed to vary. 191

A-3 Location of seven continua generated in $x'y'$ space with fixed values of $F1$ and $F2$ with $F3$ allowed to vary. Crosses indicate continua with a fixed $F1$ and $F2$ changing with each continuum. Squares indicate continua with a fixed $F2$ and $F1$ changing with each continuum. 192
A-4 "Side" view in $y'z'$ space of continua from Figure A-3. 193
A-5 Location of eight continua generated in $x'y'$ space with a fixed $F3$ and $F1$ and $F2$ maintained in constant ratios. 194
A-6 Location of continuum generated in $x'y'$ space with $F3$ fixed and increasingly greater separation in $F1$ and $F2$. 195
A-7 Location of three continua generated in $x'y'$ space parallel to the y' axis. SR, $F3$, and a constant c are fixed. 197

C-1 Spectral envelope derived from FFT of [IY] reference token. 205
C-2 Spectral envelope derived from FFT of [IH] reference token. 206
C-3 Spectral envelope derived from FFT of [EH] reference token. 207
C-4 Spectral envelope derived from FFT of [AE] reference token. 208
C-5 Spectral envelope derived from FFT of [AA] reference token. 209
C-6 Spectral envelope derived from FFT of [AO] reference token. 210
C-7 Spectral envelope derived from FFT of [AH] reference token. 211
C-8 Spectral envelope derived from FFT of [UH] reference token. 212
C-9 Spectral envelope derived from FFT of [UW] reference token. 213
C-10 Spectral envelope derived from FFT of [ER] reference token. 214
C-11 Spectral envelope derived from FFT of [IY-IH] reference token. 215
C-12 Spectral envelope derived from FFT of [IH-EH] reference token. 216
C-13 Spectral envelope derived from FFT of [EH-AE] reference token. 217
C-14 Spectral envelope derived from FFT of [AE-AH] reference token. 218
C-15 Spectral envelope derived from FFT of [AH-AA] reference token. 219
C-16 Spectral envelope derived from FFT of [AH-UH] reference token. 220
C-17 Spectral envelope derived from FFT of [UH-UW] reference token. 221

Chapter 1

Introduction and background

1.1 Introduction

A common goal of all theories of speech perception is (or should be) to explain how a listener converts the acoustic speech signal produced by a talker into a sequence of meaningful linguistic units. According to McKay (1956), theories describing the process that takes place during this conversion can be separated into two categories. Active theories view the listener as an 'active' participant in the speech-perception process and link that process with a knowledge of speech production. In these theories, the conversion is accomplished through reference to internal representations of the motor commands that produced the speech sounds encoded in the signal. Passive theories place the listener in a more passive perceptual role and consider the speech-perception process to be more sensory in nature. Here perception of the acoustic signal is generally the result of directly "decoding" it, after some amount of processing, into meaningful informational elements.

This distinction notwithstanding, Miller (1984a) suggests that all current speech-perception theories may be described by means of a generic three-stage model. The model considers only a "bottom-up" processing sequence, that is, perceiving speech without the aid of additional cues from knowledge of the language, context, talker, environment, or other "top-down" influences on speech processing. Although most theories employ top-down processing, many consider it as necessary only for ambiguous situations or instances when communication is difficult.

In stage 1, some form of spectral analysis is performed on the acoustic speech waveform, yielding information about the distribution of acoustic energy over time in terms of its frequency and intensity. In stage 2, the spectral information derived in stage 1 is transformed into perceptual dimensions. Finally, in stage 3, these perceptual dimensions are in turn transformed into linguistically significant elements, such as phonemes, syllables, or words. While most speech theorists agree that in stage 1 the ear acts as a filter bank, performing continuous short-term spectral analyses on the incoming waveform often simulated by Fourier or linear-prediction analyses, the theoretical descriptions of stages 2 and 3 are quite diverse. Active theories such as the motor theory (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman and Mattingly, 1985) and the analysis-by-synthesis theory (Stevens and Halle, 1967) suggest that the perceptual dimensions of stage 2 are articulatory gestures and, by means of special speech decoders or matching procedures, phonemes are perceptually elicited in stage 3. Some passive theorists (Fant, 1967; Morton and Broadbent, 1967) have proposed that auditory features from stage 1 are mapped onto linguistic-phonetic features in stage 2, leading to the perception of phones in stage 3. Klatt (1979) suggested another approach whereby spectral-envelope patterns derived in stage 1 are submitted to a matching procedure with stored templates of diphone sequences in stage 2. Successful matches are then subjected to further processing with word recognition achieved in stage 3.

More recently, another passive theory has been proposed by Miller (1984). The auditory-perceptual theory (*APT*) of speech perception considers that spectral information from stage 1 elicits a sensory representation by activating sensory pointers in a phonetically relevant three-dimensional perceptual space. The sensory pointers move in the space as new spectral information is processed. A perceptual response results from an integrative-predictive process based largely on the dynamics of the sensory pointers' movement in stage 2. A phonetic-linguistic representation is achieved in stage 3 through the activation of perceptual target zones (*PTZ*) in the auditory-perceptual space by the dynamics of the perceptual response. These *PTZs* correspond to the phones of a language and, upon their activation, a neural symbol code is issued. A fourth, lexical-access stage, is required in this model to find words. It is the auditory-perceptual theory of speech perception in general

and the concept and validity of perceptual target zones for vowels in particular which will provide the framework for the following dissertation.

1.2 The Auditory-Perceptual Theory (*APT*)

The fundamental concepts of the *APT* have been most recently detailed in Miller (1989). Since a basic understanding of these concepts is central to an appreciation of the work to be described, a brief description of those concepts pertaining to perceptual target zones and non-nasalized vowels follows.

1.2.1 Auditory-perceptual space (*APS*)

The auditory-perceptual space (*APS*) is defined in three dimensions, x, y, and z, where

$$\begin{aligned} x &= log(SF3/SF2), \\ y &= log(SF1/SR), \quad and \\ z &= log(SF2/SF1). \end{aligned} \tag{1.1}$$

$SF1$, $SF2$, and $SF3$ correspond to the center frequencies in Hz of the first three significant prominences in the acoustic spectra derived from the short-term spectral analysis of the incoming speech signal, commonly termed formants. SR is a low-frequency reference, called the sensory reference, based on the current talker's vocal characteristics. The sensory reference acts as a normalizing anchor point with an initial value of 168 which is shifted up or down depending on the talker's vocal characteristics. SR is often calculated by the equation,

$$SR = 168(GMF0/168)^{1/3}, \tag{1.2}$$

where $GMF0$ represents the geometric mean of the current talker's fundamental frequency.

Spectral shapes are represented in the *APS* by means of sensory pointers. The location of a sensory pointer is determined by the coordinate values derived from a short-term spectral analysis of a time-windowed segment of the incoming acoustic waveform and is continually updated with new analysis information at regular intervals. We have found that a 24-ms

window shifted in 1-ms steps appears adequate to provide the necessary information. A sensory path is created as the sensory pointer moves through the space, directed by the continually updated spectral information. For our purposes, only glottal-source spectra, i.e., where the sound source is at the glottis and *SF*1 is present, and their associated pointer (*GSSP*) will be considered. The glottal-source sensory pointer, or *GSSP*, is activated, indicating a sensory response, whenever the analysis process detects glottal-source sound above an auditory threshold.

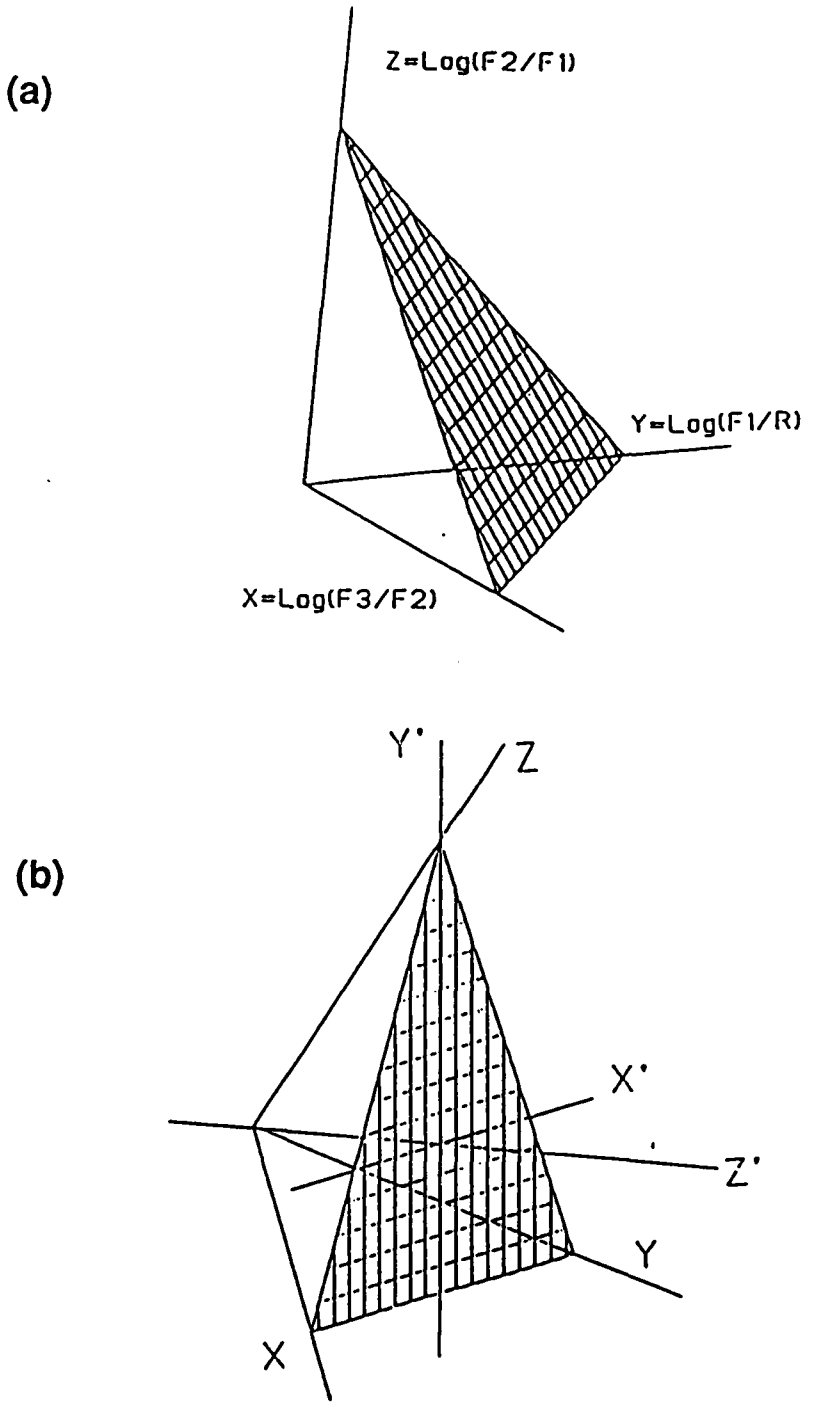
The movements by the sensory pointers are integrated into a unitary perceptual response through a sensory-perceptual transformation represented by a perceptual pointer (*PP*). The location of the perceptual pointer is controlled by the sensory pointers and a neutral point. This concept can be mathematically modeled by considering the perceptual pointer as being attached to the sensory pointers by springs such that movements by the sensory pointers result in movement of the perceptual pointer. Once again, for our purposes, only the glottal-source sensory pointer will elicit the movement of the perceptual pointer. Additionally, the perceptual pointer's movement is adjusted for changes in loudness and goodness. The perceptual pointer does not disappear when the sensory pointers are turned off, but rather, its loudness decays over 100-200 msec as it migrates to a neutral point in the *APS*.

The perceptual pointer may be continually moving through the *APS*, tracing what we call a perceptual path. Since this perceptual path reflects a continuously changing auditory experience, it is posited that the perceptual pointer performs one or a combination of several possible segmentation maneuvers to divide the continuous flow into discrete, auditory-perceptual events. The segmentation maneuvers under consideration include 1) a period of low velocity, 2) an abrupt deceleration, and 3) a high curvature in the perceptual path.

1.2.2 Formant location in the *APS*

When natural vowels are represented as points in the *APS*, the points fall within a "slab" in the *APS* often referred to as the "vowel slab," approximately located in the plane defined by $(x + y + z) = 1.18$ as shown in Figure 1-1a. As an aid to visualization, a simple rotation

Figure 1-1: (a) View of vowel "slab" (gridded area) in *APS* dimensions; (b) view of the same plane in transformed (x', y', z') dimensions.



of the *APS* axes permits a vertical orientation (Figure 1-1b) of this plane where,

$$\begin{aligned} x' &= .70711(y - x), \\ y' &= .8162(z) - .4081(x + y), \textit{and} \\ z' &= .5772(x + y + z). \end{aligned} \tag{1.3}$$

These coordinates are often referred to as "slab" coordinates and views of the *APS* utilizing these coordinates are often used for visual examination and graphical illustration of the *PTZs* for vowels and associated data.

Although the basic x, y, z coordinates of the *APS* are simple log ratios of familiar acoustic speech variables, the transformation of these coordinates to x', y', z' is somewhat more complex and makes the intuitive interpretation of formant location in terms of distance and direction in this coordinate system more difficult. In an effort to reduce this difficulty, the reader is referred to Appendix A for a graphical demonstration and discussion of how certain formant patterns manifest themselves in $x'y'z'$-space.

1.2.3 Concept and estimation of perceptual target zones (*PTZs*)

A perceptual target zone (*PTZ*) is a three-dimensional object or subspace located within the *APS*. The final stage of Miller's generic model, the perception of a phonetic representation, is thought to be accomplished through the activation of a perceptual target zone by way of a segmentation maneuver performed by the perceptual pointer within the defined boundaries of the zone. When a *PTZ* is activated, it is then said to issue a phonetic code or "neural symbol" corresponding to an allophone of the language in question. Current estimates for the shapes and locations of the perceptual target zones for the non-retroflex, monophthongal vowels of American English suggest that they have irregularly shaped, non-overlapping, and abutting boundaries.

Two approaches have been considered for estimating the locations of target zones and their boundaries with each approach having its own list of problems and advantages. In the first approach, locations of the perceptual target zones for the non-retroflex, non-diphthongized vowels of American English have been estimated on the basis of measurements of vowel productions collected from various data sources (Miller, 1987b; 1987c; Miller and

Hawks, 1986; Fourakis and Miller, 1987). These data sources include past studies from the literature as well as measurements made in our own laboratory. A front "slab" view of estimations for these zones[1] in the *APS* based on 435 such measurements can be seen in Figure 1-2, and another more recent and detailed estimation based on 2051 data points in Figure 1-3. Comparison of these figures demonstrates that as additional data points are collected, considerably more intricate boundaries seem to be required to minimize overlap between the zones. Note that the graphic visualizations in Figures 1-2 and 1-3 are two-dimensional, showing only a "front view" of the target zones by x', y' coordinates. To visualize the third dimension, a "side view" perspective is utilized with y', z' coordinates (Figure 1-4), demonstrating irregularly shaped boundaries for the earlier target zones seen in Figure 1-2 in this dimension as well. Figure 1-5 shows this perspective for the most recent target zone estimations (Figure 1-3). Note that due to technical difficulties in delineating concise boundaries in the z' dimension, these zones have been constructed with straight-line boundaries in this dimension, although in theory these boundaries are also assumed to be irregularly shaped. Thus the target zone boundaries shown in Figure 1-2 are somewhat "generalized", since, when fully detailed, these boundaries should vary relative to the z' dimension.

Certain problems exist with using measurements from natural data as the basis for zone construction. First is the problem of phonetic labeling. While all the speech data used in estimating target zones has undergone some type of phonetic verification and can be taken as having perceptual significance, it must be assumed that the vowel tokens represented from the literature were perceived as the phones they were claimed to be representative of. Various levels of phonetic verification are employed ranging from simply what the talker intended to say, to experimenter- and group-verified identifications. Instructions for phonetic identification also vary in the number of phonetic categories provided to choose from. Additional problems with this approach are found in the selection of spectra and the specification of pitch. Picking a representative spectrum for a given vowel utterance can be as arbitrary or subjective a criterion as the experimenter deems appropriate, and additional

[1]The ARPAbet symbol system (Lee and Shoup, 1980) will be used exclusively throughout this thesis for notating phonetic categories (See Table 1.1).

Table 1.1: ARPAbet symbols for representing the phoneme-like units of English within a computer. (From Lee and Shoup, 1980)

Phoneme	Computer Representation		Example	Phoneme	Computer Representation		Example
	1-Character	2-Characters			1-Character	2-Characters	
i	i	IY	beat	p	p	P	pet
ɪ	I	IH	bit	t	t	T	ten
e	e	EY	bait	k	k	K	kit
ɛ	E	EH	bet	b	b	B	bet
æ	@	AE	bat	d	d	D	debt
ɑ	a	AA	Bob	g	g	G	get
ʌ	A	AH	but	h	h	HH	hat
ɔ	c	AO	bought	f	f	F	fat
o	o	OW	boat	θ	0	TH	thing
U	U	UH	book	s	s	S	sat
u	u	UW	boot	š or ʃ	S	SH	shut
ə	x	AX	about	v	v	V	vat
ɨ	X	IX	roses	ð	D	DH	that
ɝ	R	ER	bird	z	z	Z	zoo
aU or aw	W	AW	down	ž or ʒ	Z	ZH	azure
ɔɪ or ɔy	Y	AY	buy	č	C	CH	church
ɔɪ or ɔy	O	OY	boy	ǰ	J	JH	judge
y	y	Y	you	ʍ	H	WH	which
w	w	W	wit	syl l, l̩	L	EL	battle
r	r	R	rent	syl m, m̩	M	EM	bottom
l	l	L	let	syl n, n̩	N	EN	button
m	m	M	met	flapped t,ɾ	F	DX	batter
n	n	N	net	glottal stop, ʔ	Q	Q	
ŋ	G	NX	sing	Silence	-	-	
				non-speech Segment	!	!	laugh, etc.

AUXILIARY SYMBOLS (1- AND 2-CHARACTER CODES ARE IDENTICAL)

Symbol	Meaning	Symbol	Meaning
+	Morpheme boundary	:3 or .	Fall-rise or non-term juncture
/	Word boundary	* **	Comment (anything except * or **)
#	Utterance boundary	' '	Apos.-surround special symbol in comment
:	Tone group boundary	( )	Phoneme class information
:1 or .	Falling or decl. juncture	< >	Phonetic or allophonic escape
:2 or ?	Rising or inter. juncture		

STRESS REPRESENTATIONS (IF PRESENT, MUST IMMEDIATELY FOLLOW THE VOWEL)

Value	Stress Assignment	Value	Stress Assignment
0	No stress	3	Tertiary stress
1	Primary stress	ˈ	(Etc.)
2	Secondary Stress	ˌ	

Figure 1-2: Estimations of perceptual target zones for American English in APS $x'y'$ coordinates based on measurements of 435 vowels from natural speech.

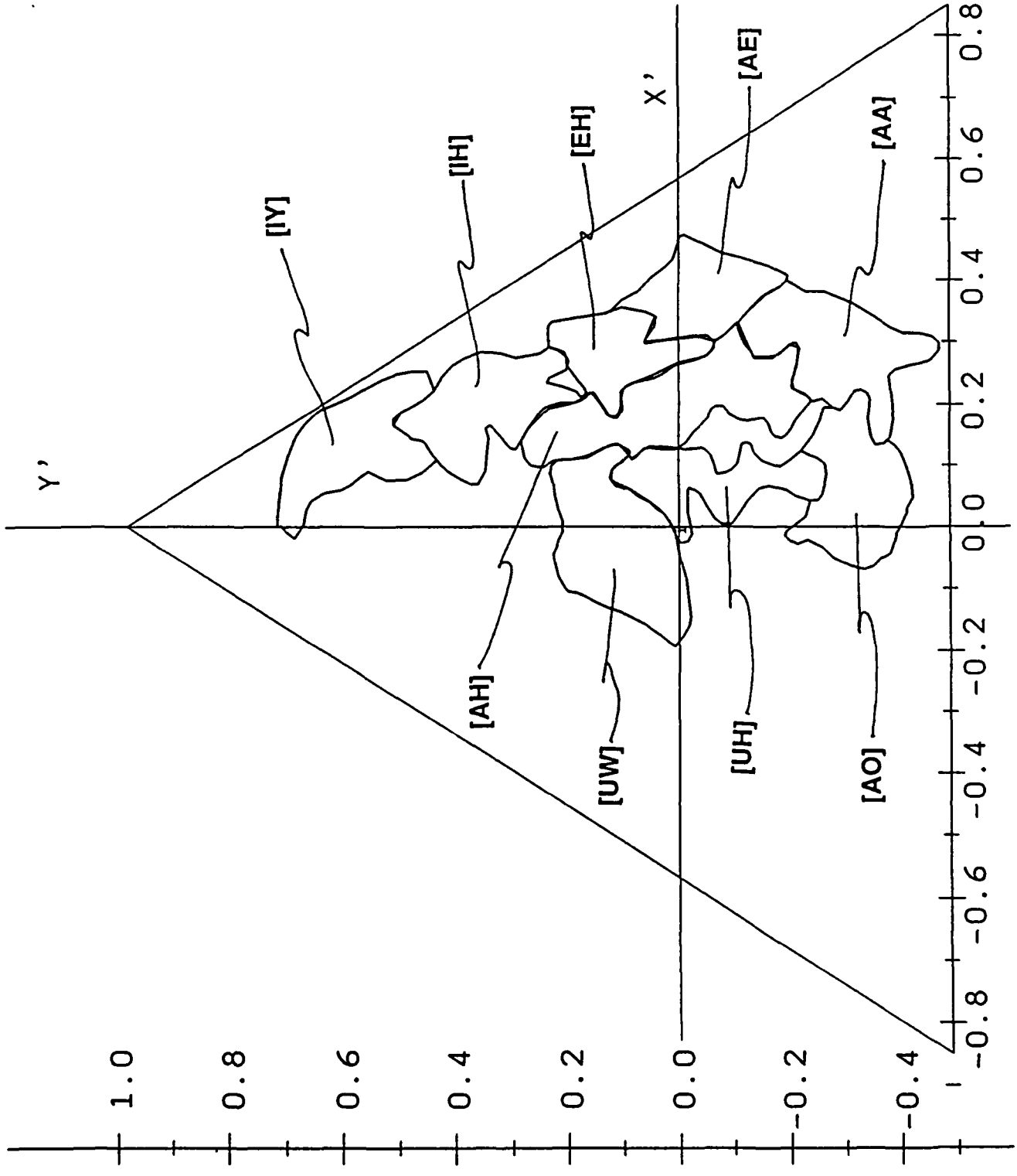


Figure 1-3: Estimations of perceptual target zones for American English in APS $x'y'$ coordinates based on measurements of 2051 vowels from natural speech.

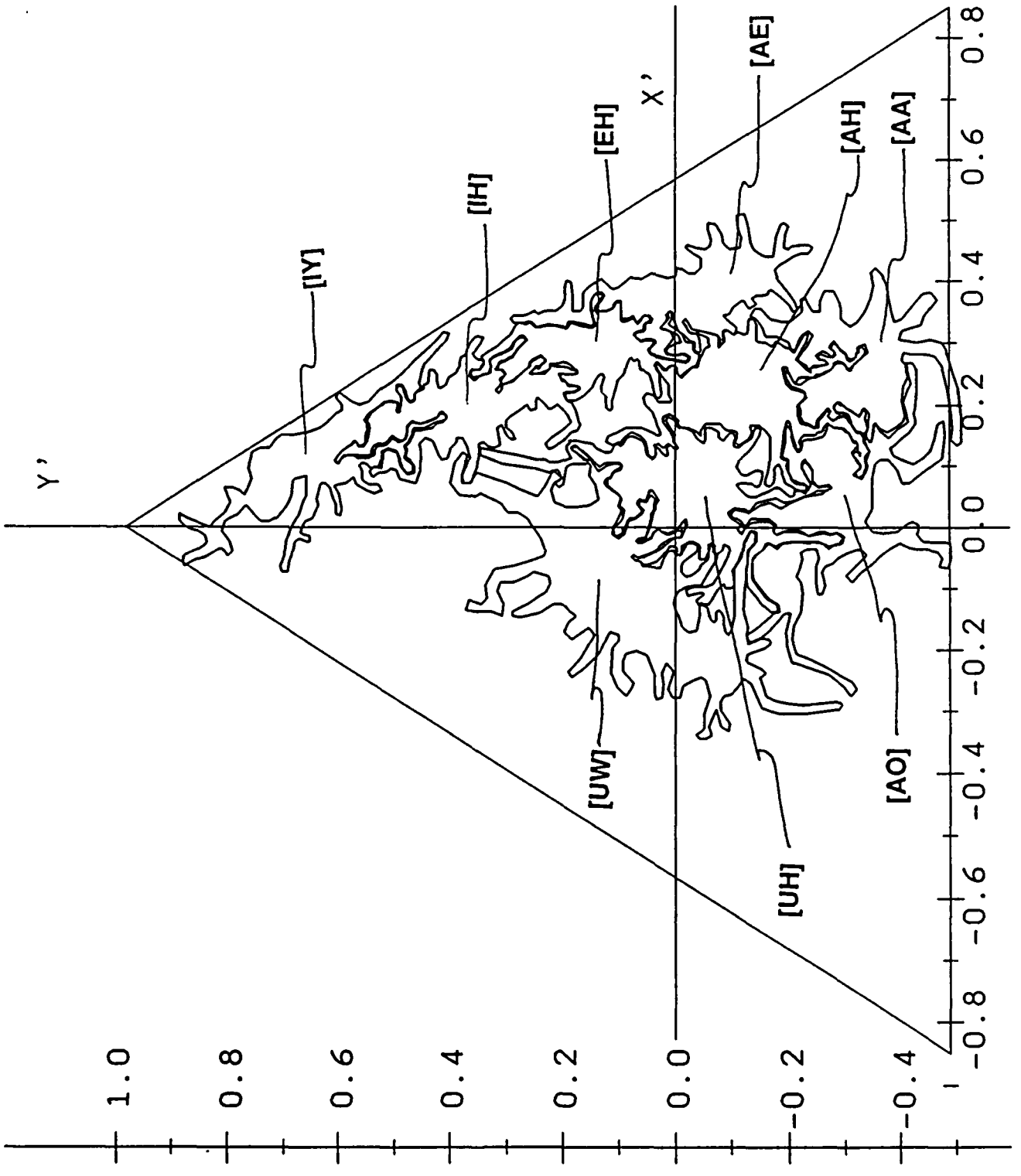


Figure 1-4: Estimations of perceptual target zones for American English in APS $y'z'$ coordinates based on measurements of 435 vowels from natural speech.

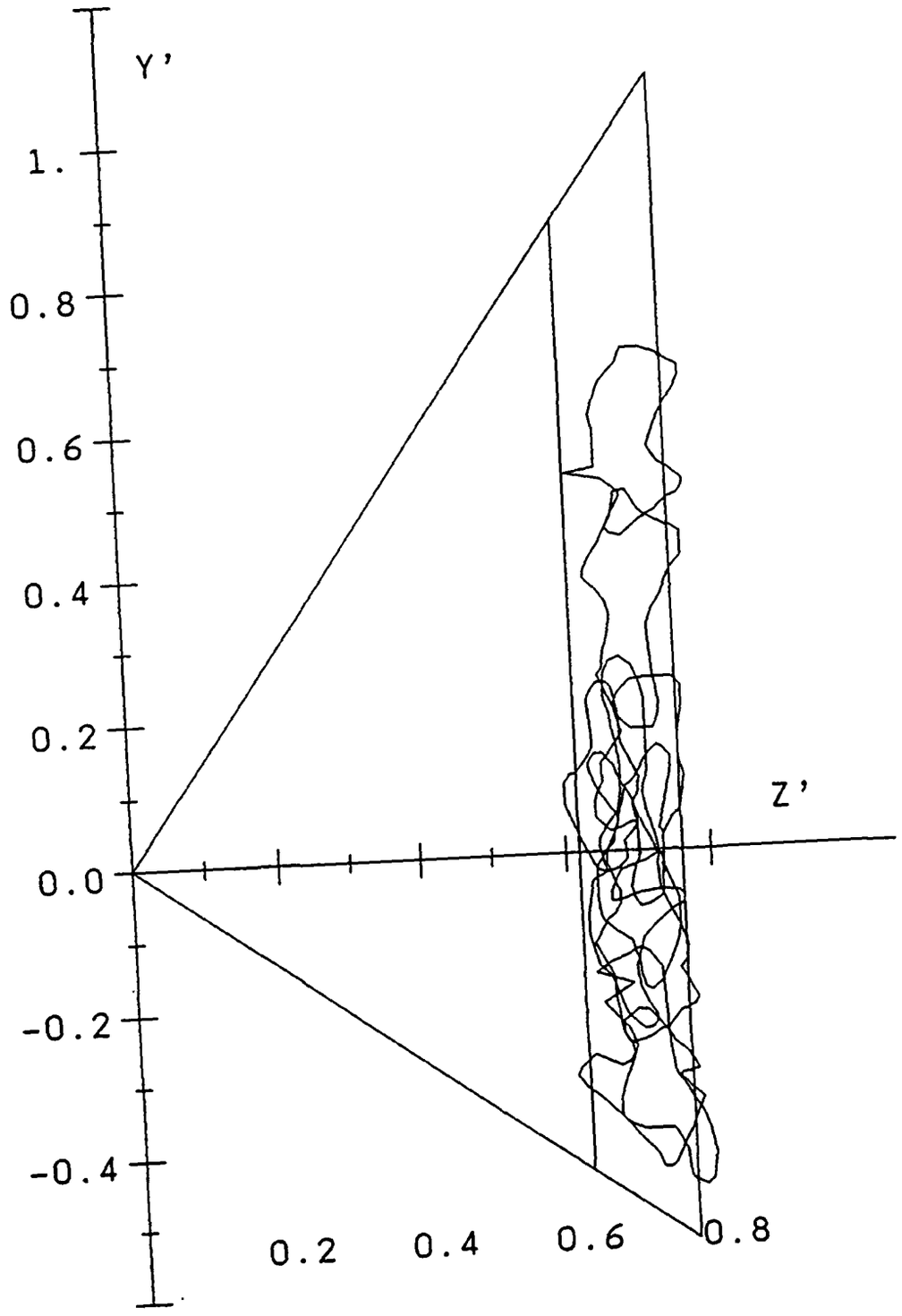
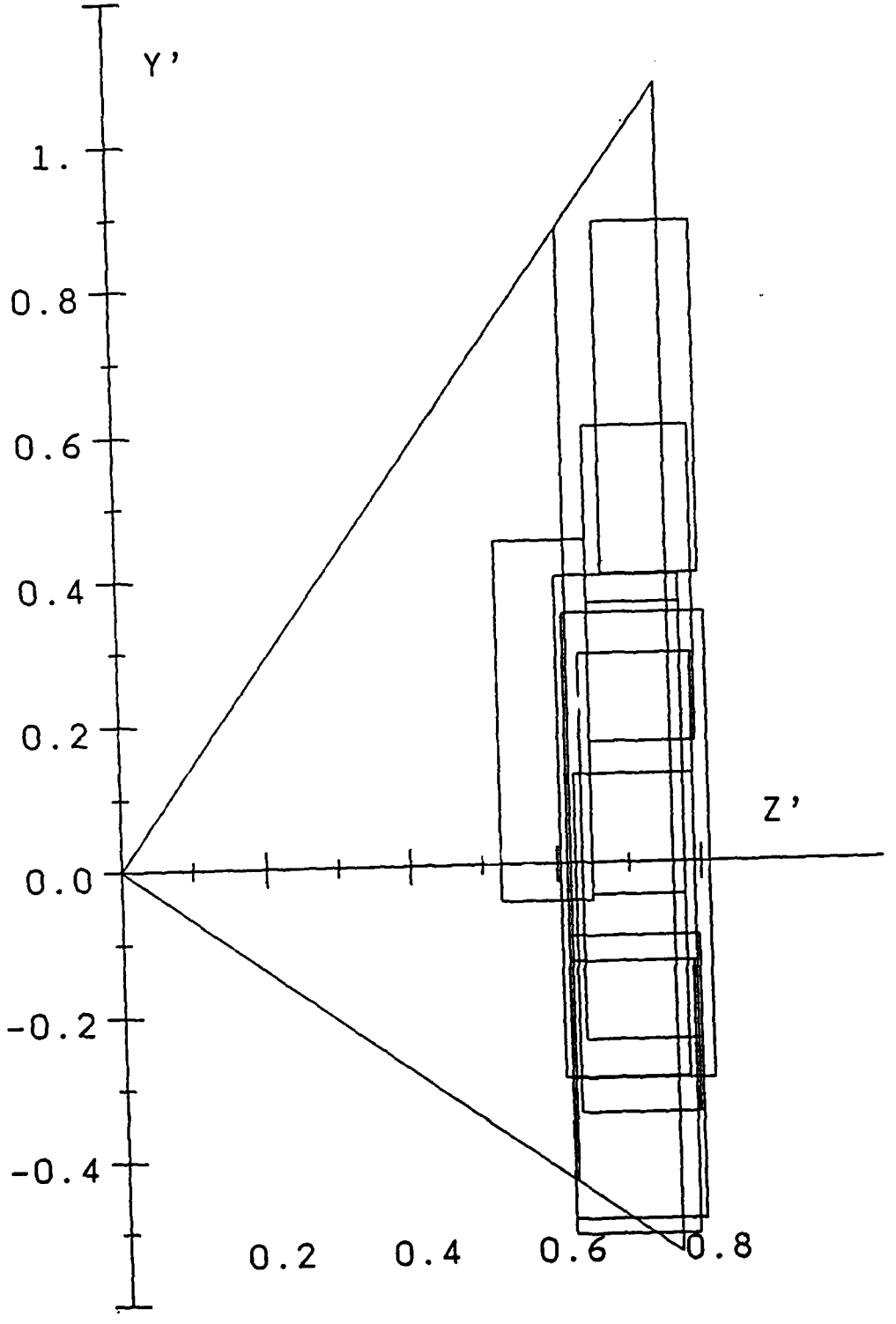


Figure 1-5: Estimations of perceptual target zones for American English in APS $y'z'$ coordinates based on measurements of 2051 vowels from natural speech.



variance can stem from the different measurement techniques which have been employed through the years.

It is currently hypothesized that, given a sufficient number of data points gathered from more talkers and more vowel contexts, the *PTZ* boundaries will stabilize, although there is neither guarantee that this will happen nor an estimation as to the number of points, talkers, or contexts this stabilization may require. This uncertainty questions the validity of basing perceptual target zones solely on such an approach.

The second approach to determining the locations of perceptual target zones utilizes the phonetic identifications of synthetic speech as a basis for mapping the *APS* vowel space. The speech synthesizer is a powerful and often-used scientific tool for investigating hypotheses about complex sounds and has proven particularly useful in speech and auditory research. The synthetic-speech approach has the advantage that, unlike the natural- speech approach which must rely on chance for phonetic identification of locations in *APS*, specific locations in the *APS* can be synthesized and phonetically identified, providing a more complete view of perceptual target zones and their boundaries.

However, the synthetic-speech approach is not without its problems. Since all the variations for the parameters comprising natural speech that elicit perceptual differences in vowel quality have yet to be specified, it is impossible to synthesize vowels which sound completely natural. Furthermore, while it is safe to assume that all these parameters are present in the vowel utterances utilized in the natural speech approach, it can only be speculated that the physical attributes necessary for perceptual salience of synthetic vowels are present. Additionally, the parameters selected to be varied must be questioned as to whether they are correct and sufficient for changing perceived vowel quality. Thus, the validity of using this approach alone is also questioned.

Given consideration of these two approaches for estimation of perceptual target zones and their boundaries, it becomes apparent that neither is clearly superior and that the best solution may be to utilize both approaches with the hope that the results will converge and verify one another.

1.3 Overview of experiments

For decades speech researchers have attempted to find a method or metric based on acoustic parameters whereby the vowels of a language could be described with unique and independent categories, despite differences in talkers, speaking rate, stress, and context. The concept of the perceptual target zone offers a potential answer to this question and provides as well a major underlying foundation to the auditory-perceptual theory. Thus, the validation of this concept is critical not only to the ongoing research efforts within the theory's present framework, but also to the general acceptance of such a theory by the scientific community.

Two experiments will be outlined here aimed at exploring the validity of perceptual target zones for vowels. Experiment I will investigate, by means of phonetic identifications and confidence ratings, listeners' perceptions of simple isolated vowel sounds which have been synthesized from variables provided by distinct points in the *APS*. The results of this experiment will provide a phonetic map of the areas in and around the vowel slab and offer a basis for comparison between this and other vowel classification approaches, including the current *PTZ* estimates. The results of Experiment I will also provide information about individual perceptual differences and mapping reliability and serve as guidelines for Experiment II.

Experiment II will investigate the *APS* at very fine levels of resolution. Indeed, the goal of Experiment II is to estimate the difference limen for various locations in the *APS*. An adaptive up-down procedure utilizing a cued, two-alternative, forced-choice (2AFC) task will be employed. The results of this experiment provide information about the resolution necessary for exploring more precise *PTZ* boundaries, as well as address questions concerning the potential differences in discrimination sensitivity within and between *PTZs*. Additionally, evidence will be presented demonstrating how discrimination of vowel sounds is affected by multiple simultaneous formant changes.

Chapter 2

Experiment I: Perceptual Mapping of the APS Vowel Space.

2.1 Introduction

The purpose of Experiment I is to gather data which can be used to estimate the sizes, shapes, and locations of target zones for the vowels of American English in the *APS* and additionally, to attempt to validate the concept that such zones are non-overlapping, enabling their use for vowel classification. The target zone estimates will be made by means of listeners' identifications, and corresponding confidence ratings, of synthetic tokens which represent specific locations in the auditory-perceptual space. The experiment is motivated by the fact that inspection of estimates for these zones based on the presently available production data (Figures 1-9 and 1-10) shows that portions of the space in the vowel slab remain unaccounted for. Several reasons may account for this. 1) As was previously discussed, the current database may be too limited in terms of an adequate sample of subjects and phonetic environments to modify the current boundaries further. This may be particularly true for unaccounted-for spaces between the *PTZs* and the exterior edges of boundaries. 2) Vocalic sounds corresponding to points in the unaccounted-for spaces may represent speech sounds or portions of speech sounds other than vowels. There is reason to believe that the *PTZs* for [L,R,W,Y] may occupy portions of the vowel slab, as well as the voiced portions of the fricatives [Z,V,DH,ZH,JH], flaps [DX], and voice bars (Miller and Hawks, 1986). 3) Vocalic

sounds corresponding to points in the unaccounted-for spaces may represent speech sounds from languages other than American English. In particular, rounded front vowels, such as those found in German and Swedish may uniquely occupy some of the unaccounted-for space (Jongman, Fourakis, and Sereno, 1989). 4) Vocalic sounds corresponding to points in the unaccounted-for spaces may be perceived as belonging to specific phoneme categories, but are not generally realizable by the human vocal apparatus. Preliminary work indicates that some synthetic vowel sounds utilizing extreme formant values or unlikely relations between formant values can be perceived as English vowels. 5) Vocalic sounds corresponding to points in the unaccounted-for spaces may sound completely unspeechlike. Computation of the values of $SF1$, $SF2$, and $SF3$ for a given value of SR in the spaces currently unaccounted for reveal that some of these points may not be physiologically realizable by the human vocal tract, but should be physically realizable with a digital formant synthesizer (i.e., $F0 < F1 < F2 < F3$). Given the primary purposes and exploratory nature of this experiment, all of these possible reasons cannot be addressed and thus become part of the goals for future research. Delineation of non-vowels will not be attempted and identification of vowels from languages other than American English will be dealt with in only a qualitative manner. Some attempt will be made, however, to limit tokens to those which potentially may be produced by a male voal tract.

Experiment I is additionally motivated by the fact that no studies to our knowledge have investigated an extensive range of sounds that humans are able to produce or perceive as vowels. Many perceptual studies of vowel spaces and their boundaries have been based on the identification of natural speech (Potter and Peterson, 1948; Peterson and Barney, 1952; Fairbanks and Grubb, 1961; Pols, van der Kamp, and Plomp, 1969), and in these experiments an exact description of the stimuli is often difficult to obtain. Of the studies dealing with the perception of synthetic stimuli, most have utilized straight-line (relative to an F1-F2 plot) continua which cross through several vowel spaces (Fry, Abramson, Eimas, and Liberman, 1962; Stevens, Liberman, Studdert-Kennedy, and Öhman, 1969; Repp, Healey, and Crowder, 1979).

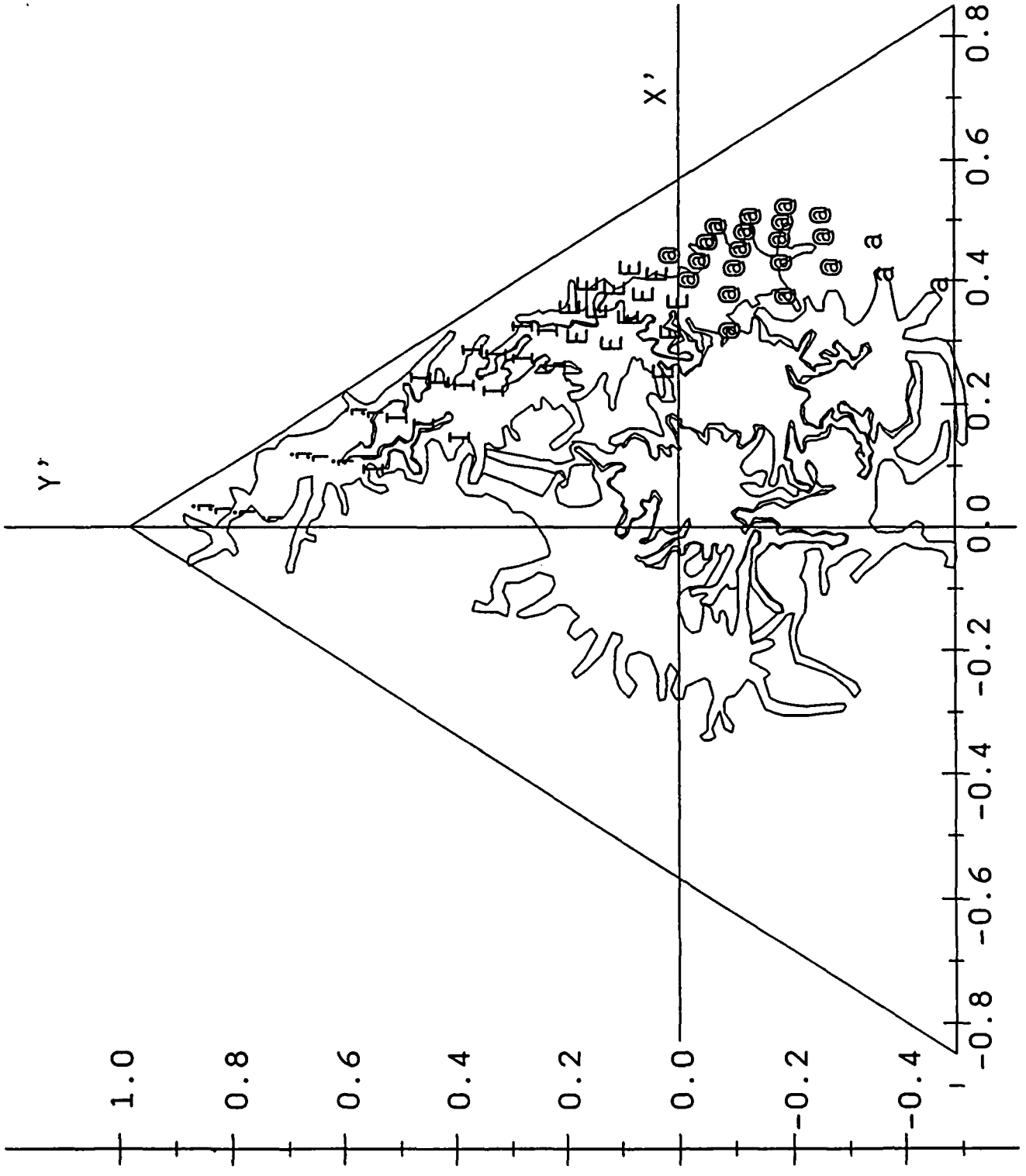
Only a few studies have utilized a relatively wide range of possible F1-F2 combinations (Holmes, 1986; Scholes, 1967; Ainsworth and Millar, 1971; Millar and Ainsworth, 1972;

Nearey, 1977; R.L. Miller, 1953). Of these, only R.L. Miller (1953) and J. Holmes (1986) included a variable F3 for some stimuli. Unfortunately, subjects in the Holmes (1986) study were instructed to identify only tokens judged to be exactly phonetically correct and thus other additional information relevant to the vowel categories and their boundaries was lost. However, in the R.L. Miller (1953) study, several hundred two- and three-formant vowel tokens were synthesized with a man's (144 Hz) and child's (288 Hz) fundamental frequency and presented to subjects for identification. Among the findings were that (1) general areas assigned to vowels as plotted in an $F1 - F2$ space remained relatively fixed for sounds of different fundamental frequencies; (2) the addition of a third formant added considerably to the unanimity of responses, particularly for front vowels; and (3) the presence or absence of sounds from certain regions may influence a subject's response to sounds of other regions. Visual inspection of the $F1 - F2$ plots from this study suggests that (1) boundaries between vowels may be abutting and irregularly shaped, and (2) these boundaries may be quite narrow. The nature of these boundaries however, was not the focus of this study. Results from this study of identifications for stimuli utilizing a third formant have been plotted as points in the auditory-perceptual space (Figure 2-1) and approximately 85% of the identifications are found to be either in agreement with the current *PTZ* estimates or lie in adjacent unclaimed space.

Given these motivations, several categories of questions may be addressed. The first category pertains to the perceptual responses themselves. Given that all tokens can be identified as vowels of American English, what acoustic variables of the stimuli mediate these responses? How well do subjects agree on identifications and confidence ratings and how reliable are these agreements? Are identification agreements and confidence ratings correlated or do they reflect different kinds of information about the stimuli? What role do individual differences play in the results?

A second category of questions is directed toward the concept of vowel zones in the *APS*. Can zones for the vowels of American English be constructed on the basis of the results of such an experiment? If so, what are the locations and shapes of the zones when all synthetically realizable space in the *APS* is considered, and how do they compare to the *PTZs* based on natural speech? What is the nature of the boundaries between the zones,

Figure 2-1: Subjects' identifications for synthetic vowels from R.L. Miller (1953) plotted in APS $x'y'$ coordinates along with current target zone estimates from Figure 1-3.



that is, how well defined are they, and how do they vary between listeners? Can a measure of vowel "saliency" be derived from the responses, and if so, how is saliency distributed within and between the zones?

The last category addresses the concept of target zones in the *APS* as a classification scheme for vowels. How do target zones based on the results of this experiment compare to other vowel classification schemes in their ability to correctly classify synthetic and natural vowels ?

2.2 Methods

2.2.1 Stimuli

Isolated vowel sounds representing points in the *APS* were synthesized with 16-bit resolution at a 10 kHz sampling rate using the cascade portion of a Klatt digital formant synthesizer (Klatt, 1980), implemented on a DEC MicroVax II computer. These points are equi-distant .05 log unit (whole tone) steps in the $x'y'$-plane and arbitrarily originate from x'= 0.0, y'= 0.0, such that all coordinate values in both dimensions are evenly divisible by 0.05. Figure 2-2 shows the points for one z' slab ($z' = 0.70$). The points span seven slabs in the $z'y'$-plane (Figure 2-3), ranging from $z' = 0.50$ to 0.80 in 0.05 log unit steps. This range is based on the z' values for data points in the CID natural speech database[1] and the Peterson and Barney (1952) study where the predominant z' range is 0.65 to 0.75 for non-retroflex vowels. However, to include consideration of the retroflex /ER/, the range of z' must extend back to at least $z' = 0.53$ in accordance with the locations of natural data points. Thus the range of $z' = 0.50$ to 0.80 ensures that the range of vowel sounds most likely to be produced in natural speech will be included among the synthesized stimuli. However, the three planes where $z' = 0.65$, 0.70 and 0.75 will be referred to as the *primary* planes and the planes where $z' = 0.50$, 0.55, 0.60, and 0.80 the *secondary* planes.

The 0.05 log unit resolution is thought to be poorer than that required to adequately determine *PTZ* boundaries, but sufficient to provide an indication of boundary areas between *PTZs*, as well as the general locations of the *PTZs* with a reasonable number of stimuli.

[1]For a detailed description of this database, see Miller, 1989, Appendix B

Figure 2-2: Location in APS $x'y'$ coordinates of synthesizable tokens for one z' plane ($z' = 0.700$).

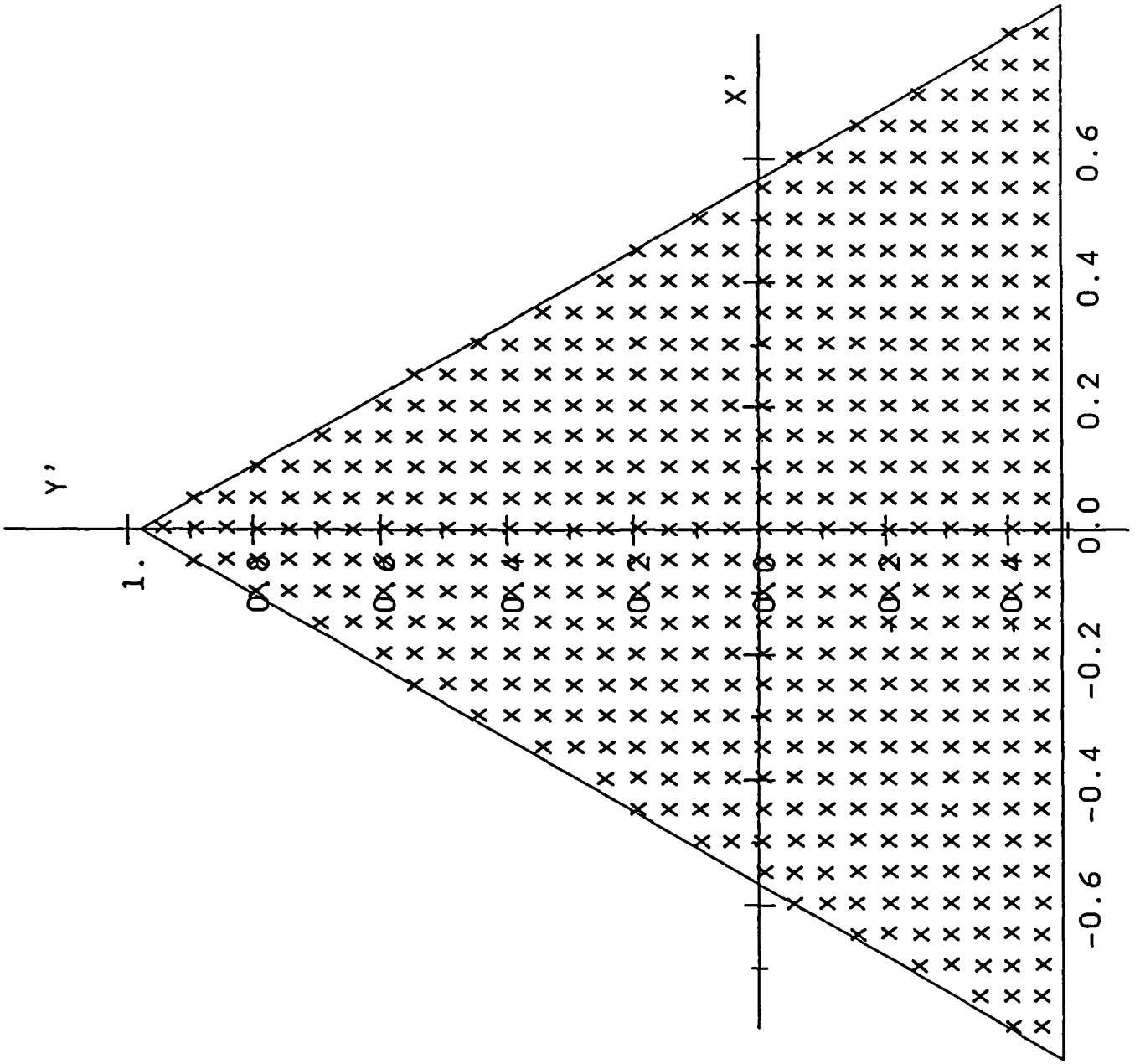
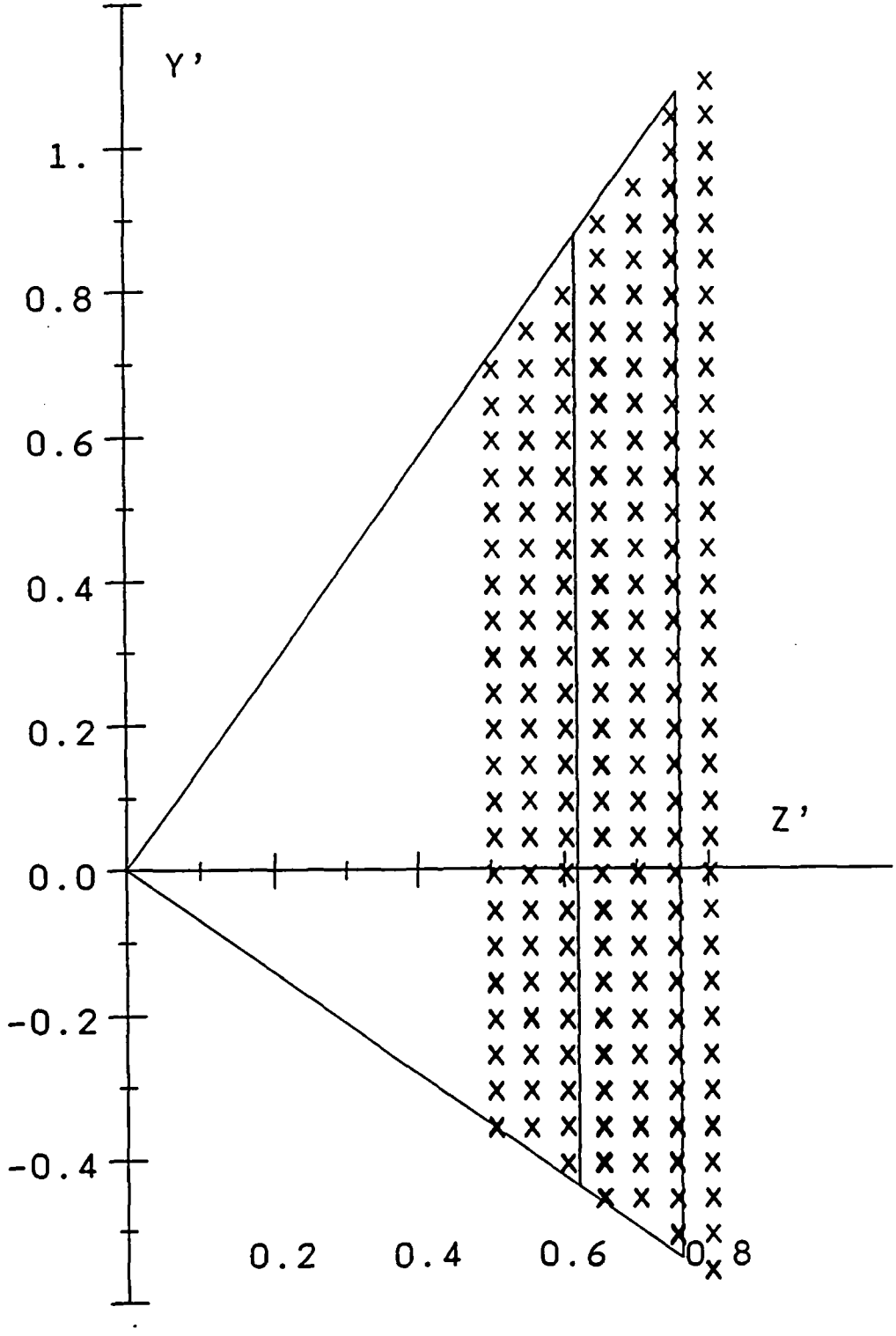


Figure 2-3: Location in APS $y'z'$ coordinates of z' planes utilized for Experiment I.



Valid points were initially considered to be those where $SR \leq F1 \leq F2 \leq F3$. This criterion yields 3151 possible points for synthesis. However, results from pilot studies indicated that a number of these synthetic tokens, particularly those near the outer borders of the vowel space, sound very unnatural or "unspeechlike." To eliminate these tokens and to reduce the total possible tokens to a more manageable number, a second set of criteria was employed. Rules based on the CID natural-speech-data corpus for specifying the possible frequency-band ranges of $F1, F2$, and $F3$ relative to SR have been developed for possible use in an automatic formant-picking procedure. These formant-band range (FBR) estimates indicate the minimum and maximum allowable values of $log(Fn/SR)$, where n = 1, 2 or 3. The current estimates of these values[2] are,

$$\begin{aligned} 0.09 &\leq log(F1/SR) \leq 0.85 \\ 0.56 &\leq log(F2/SR) \leq 1.20 \\ 0.86 &\leq log(F3/SR) \leq 1.40. \end{aligned} \tag{2.1}$$

By using these rules as additional criteria, the total number of acceptable tokens was reduced by 45% to 1725. Figure 2-4 shows the acceptable points in the same z' slab as in Figure 2-2 plotted over the most current estimates for PTZ boundaries from Figure 1-3 in Chapter 1. Note that virtually all the space occupied by the $PTZs$ is acceptable for synthesis.

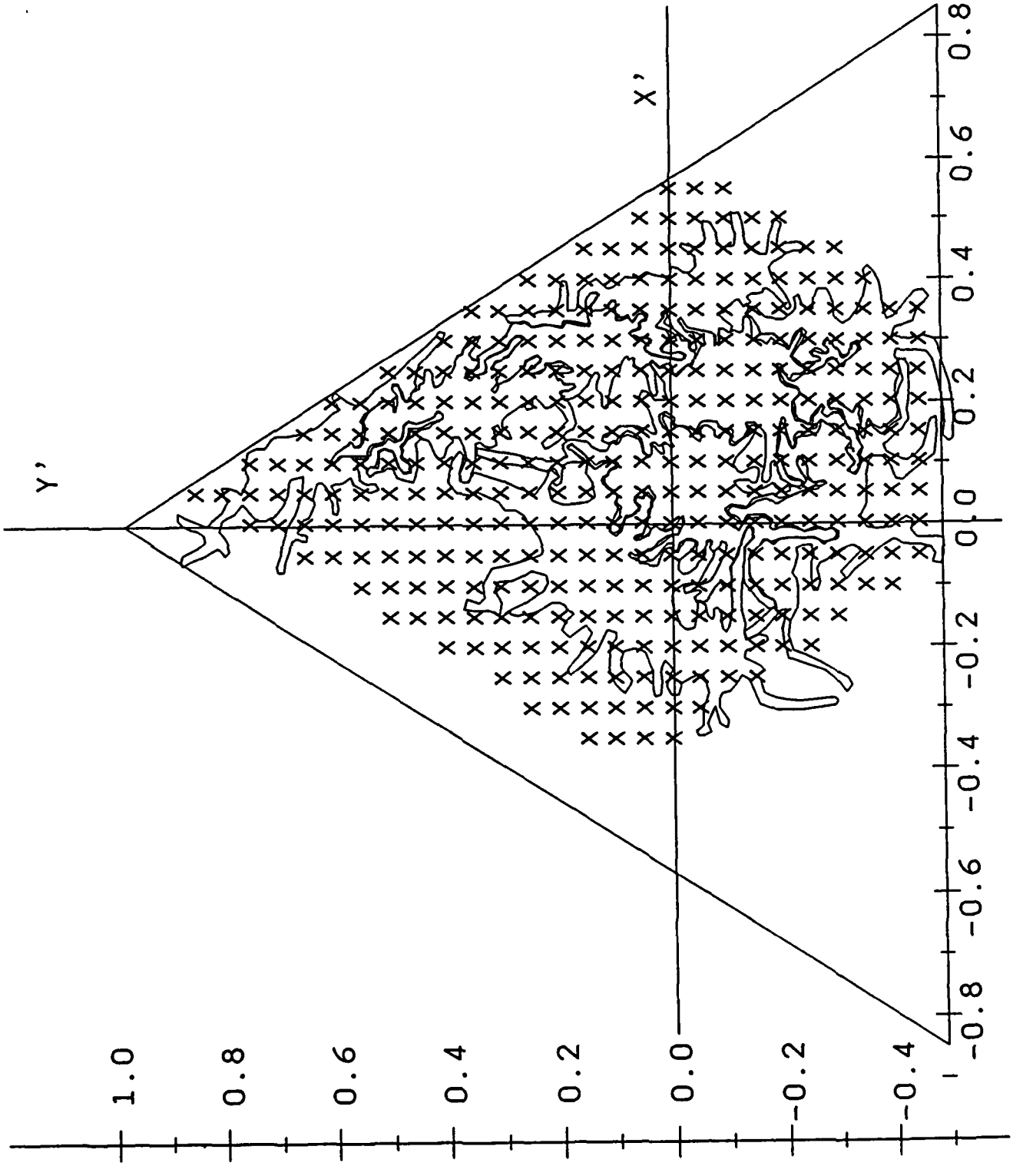
Frequency values for $F1, F2$, and $F3$ were calculated from each set of x', y', z' coordinates by means of a matrix equation. The value of the sensory reference (SR) was held constant at 155 for all tokens, yielding a fundamental frequency ($F0$) of 132 Hz and FBR ranges of

$$\begin{aligned} 190 < F1 < 1097 \\ 563 < F2 < 2456 \\ 1122 < F3 < 3890. \end{aligned} \tag{2.2}$$

$F4$ is arbitrarily set to 4000 Hz, somewhat higher than would be typically found for a male talker, to accomodate the use of higher-than-normal values of $F3$. All tokens were 400 ms in duration[3]. The amplitude contour of each token was ramped by logarithmically

[2]The minimum value of the FBR for $F1$ is normally set to 0.00 to provide for instances where $F1$ may drop to the value of SR as in voice bars. However, for vowels, $F1$ does not fall below 190 Hz for any data found in the CID data corpus. Thus, the minimum value of the FBR for $F1$ has been set to 0.09 ($log(190/155)$) for the purposes of this experiment.

Figure 2-4: Locations in $x'y'$ of tokens in one z' plane ($z' = 0.700$) acceptable for synthesis after applying formant-range limiting criteria plotted with the most recent target zone estimations from Figure 1-3.



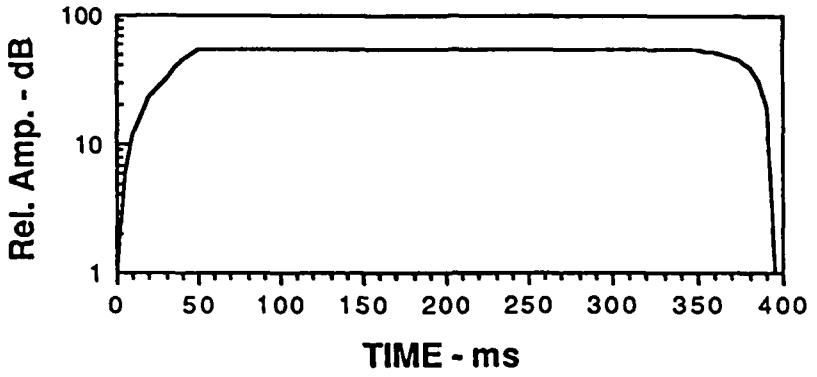
interpolating from 1 to 55 dB over the first 50 ms, held at a constant 55 dB for the next 300 ms, and again ramped by log interpolation from 55 to 1 dB over the last 50 ms (Figure 2-5a). The $F0$ contour of each token followed approximately a scaled version of the parameters used by Burdick and Miller (1975), linearly interpolating from 114 to 132 Hz over the first 50 ms, maintaining a steady-state 132 Hz over the next 150 ms, and again linearly interpolating from 132 to 100 Hz over the last 200 ms (Figure 2-5b). Amplitude normalization across tokens was achieved by means of scaling all tokens to equal peak amplitude.

Values for the first three formant bandwidths ($B1, B2, B3$) were generated by means of an equation based on data from Miller (1980). In this study, bandwidth (BW) estimates for males from Dunn (1961) and Fujimura and Lindqvist (1971) were averaged and plotted as Q (Fc/BW) over the ratio of the formant center frequency (Fc) to the fundamental frequency ($Fc/F0$). These data were then fit with a curve by means of a 5th-order polynomial regression (See Figure 2-5c). The coefficients from this regression are used to calculate the formant bandwidths. The bandwidth values generated by this equation agree reasonably well with other published values and predictive formulas (House and Stevens, 1958; Fant, 1972). All other synthesis parameter specifications are listed in Appendix B. Examples of spectral envelopes from the stimuli are shown in Appendix C.

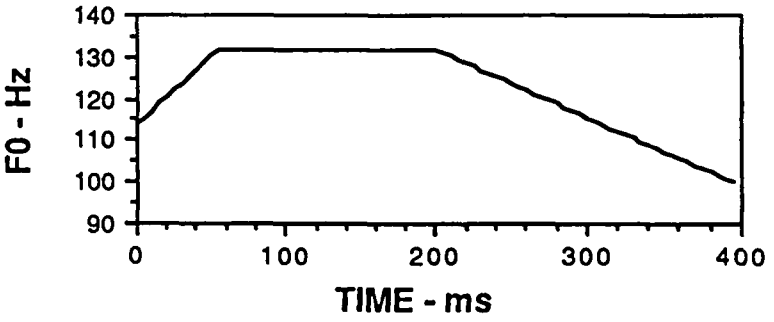
[3]The lax vowels of English [IH, EH, AE, AH, UH], with the exception of [AE] are often considered the short vowels and their duration in natural speech averages about 80% of the vowels considered intrinsically long in English, [IY, AA, AO, UW], even when spoken in isolation (Strange, Edman, and Jenkins, 1979). The differences in vowel duration have long been considered a "secondary cue" to vowel perception (Peterson and Lehiste, 1960). However, past studies addressing this issue have reported conflicting results. Ainsworth (1972) demonstrated that duration can effect the identification of synthetic vowels. This result must be viewed in light of the fact that two-formant vowels were utilized in this study which may have sounded less natural and perceptually less salient than natural speech, therefore potentially increasing the weight given to durational cues as a response criterion. Additionally, Ainsworth (1972) suggests that the importance of duration as a cue may be less important to isolated vowels than vowels in a linguistic context. Pisoni (1971), Stevens, et al. (1969), and Repp, et al., (1979) all found that the [IH] category identification responses to be less stable than the [IY] or [EH] responses and speculated that this instability could be due to the relatively long stimulus durations compared to natural speech. However, the current experiment is not as concerned with response stability within vowel categories as it is with the stability of boundaries between categories. The three studies previously mentioned all found relatively sharp boundaries between categories at approximately the same place in the vowel stimuli continuum. Additionally, Pisoni(1973) demonstrated that this same boundary (for IY-IH) sharpness and location remains stable with synthetic vowels of both long (300 ms) and short (50 ms) durations.

Figure 2-5: (a) Overall amplitude contour used for token synthesis; (b) Fundamental frequency ($F0$) contour used for token synthesis; (c) Q as a function of the ratio of formant center frequency (Fc) over fundamental frequency ($F0$) used for formant bandwidth calculation in token synthesis (See text).

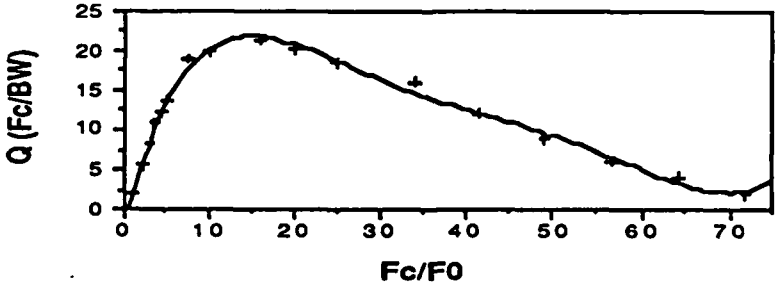
(a)



(b)



(c)



2.2.2 Procedure

Identification and confidence-rating tasks were employed for the mapping of the *APS* vowel slab. The tokens were randomized, low-pass filtered at 5 kHz, and presented via a digital-to-analog converter (MicroTechnology Unlimited DigiSound-16) directly from the computer. Subjects listened binaurally by headphone (AKG K-141) in a sound-attenuated room at a comfortable listening level ($\approx$ 55-60 dBA-Slow SPL). Subjects were asked to identify each token as one of the following vowels, [IY, IH, EH, EY, AE, AA, AH, AO, OW, UH, UW, ER]. Although [EY] and [OW] are generally considered diphthongal nuclei and not pure vowels in American English, they are included here to allow more thorough comparison with previous mapping experiments (Miller, 1953) and because they may be considered phonetically as single target speech sounds in American English (Lehiste and Peterson, 1961). Currently no target zones exist for [EY] and [OW], so estimations of these zones will be required.

In an earlier pilot study, subjects were asked to rate each token for its "goodness" on a 9-point scale. However, these subjects indicated that the nine-point goodness-rating scale was too broad to establish and monitor, suggesting that a rating scale using fewer categories should be considered. Furthermore, it was determined that the "goodness" rating reflected too many different variables, so the rating criteria were reduced to reflect only the certainty of the response. Instead of a goodness rating, subjects were asked to rate each token identification response for "confidence" on a 5-point scale. Subjects were instructed that the confidence rating should reflect how certain they were as to their identification of the token. A rating of 1 should reflect that the subject was very unsure as to the category they selected, a rating of 2, 3, or 4 should reflect moderate levels of certainty, and a rating of 5 should reflect that the subject was very confident of their response.

The 1725 tokens were randomly divided into 17 blocks with 16 blocks of 100 tokens each and 1 block containing 125 tokens. The presentation order of the blocks was randomized per subject as well. Sessions were under computer control and structured in the following manner. A row of 12 boxes appeared on the terminal screen. Within each box was a two-character arpabet symbol for one of the categories and a [hVd] word containing that vowel sound. A message flashed on the screen that the token would be presented in two seconds. After the token was presented, the subject was prompted for a response. The identification

responses were made by pressing keys labeled with the same two-character symbols. After the identification response was entered, the subject was prompted for a confidence rating response. A row of boxes containing the numbers 1 through 5 with box 1 labeled "very unsure," box 3 labeled "mediocre," and box 5 labeled "very sure" appeared and the subject selected their response by numerical key entry. This concluded a single token presentation. Subjects had the option of repeating the token being judged a number of times if necessary at any point in the process prior to giving the rating response. The number of times a token was repeated was also recorded. Subjects kept a written account of errors they had made in key entry which the experimenter later corrected. Each subject evaluated all 1725 tokens twice.

2.2.3 Subjects

Subjects, five male and five female, were recruited from the student body of Washington University and the nearby St. Louis area. Subjects' ages ranged from 17 to 25 years. All subjects were native speakers of American English with no known history of either speech or hearing impairment. All subjects were naive in terms of formal phonetic training. Six of the subjects were born and raised in the Midwest and the remaining four had resided in St. Louis a minimum of two years. While some dialectal differences between subjects were anticipated, all subjects were screened to ensure that he or she normally used and perceived in their everyday speech all of the vowel sounds to be used as response categories in the experiment. Two subjects, one male and one female, were unable to complete the experiment, thus the results presented here will reflect the data collected from eight subjects.

Training

Subjects were trained in the experimental protocol in two stages. In the first stage, subjects trained on 51 tokens consisting of: 1) synthetic vowels constructed using the male average formant frequencies values from Peterson and Barney (1952); 2) exemplar tokens selected from the test stimuli; and 3) vowels spoken and recorded by the experimenter imitating the test stimuli. Subjects identified ten randomizations of these tokens or achieved a consistent identification rate of at least 96% correct, whichever came first.

In the second stage of training, subjects identified and provided confidence ratings for the 304 test tokens from the $z' = 0.70$ plane later in the main experiment. This plane lies midway along the z' dimension in the vowel slab and vowel tokens synthesized from this plane generally evoke salient perceptions of eleven of the twelve categories[4] used for identification. The 304 tokens were randomized and divided into three blocks. Subjects' identifications were analyzed to ensure that all eleven categories were used and that there was a general consistency of grouping for like identifications . The results of this informal analysis were discussed with the subject. One subject (1M) did not receive this portion of the training. This concluded the second stage of training.

2.3 Results

2.3.1 General observations

A number of observations were noted from conversations with subjects following the experiment which are of interest here. Subjects noted that a number of tokens did not sound like vowels of English, but rather, like the front rounded vowel /ü/ used in German, and that their perception of these tokens was ambiguous between /IY/ and /UW/. Examination of the data and informal listening by trained phoneticians indicated that these tokens fell in the area at the border of /IY/ and /UW/. Thus, a region along this border most probably should be classified as "not a vowel of English."

Although the diphthong /EY/ can be produced as a monophthong in American English (Pike, 1947; Lehiste and Peterson, 1961) and occurs as a pure vowel in many languages, several subjects noted that they had difficulty in distinguishing /EY/ from /IH/ and /EH/. This difficulty might have been reduced had more rigorous training in identifying this category been employed. Monophthongal versions of /OW/ did not seem to present such a problem.

Subjects also noted that a number of tokens sounded like something between a purely monophthongal vowel and /ER/, making these tokens difficult to classify. Informal listening by the experimenter indicated that tokens falling near the boundaries for /ER/ tended to

[4]The vowel /ER/ does not occur in this z' plane.

sound very rhotacized, or "r-colored", presumably creating the classification difficulty.

Before the analysis of the data in perceptual terms, it is desirable to know how subjects utilized the identification and confidence rating categories. Specifically, do the subjects represent a homogeneous group in terms of the identification and confidence rating responses? Are subjects' responses reliable in terms of consistency and repeatability?

2.3.2 Identifications

The frequency of responses for each identification category are shown in Table 2.1 for each subject's response sets. The subject response sets are designated by first the subject number, followed by M or F (for male or female), and then the set number (1st or 2nd replication). These designations will be referred to throughout the results section. The range of response frequencies is great, varying from as few responses as 9 (3M1-/EY/) to as many as 513 (1F1-/UW/). Category /EY/ had the fewest average number of responses overall and /UW/ the greatest with the remainder of response categories falling roughly into two groups, /IH,EH,AE,ER/ and /IY,AA,AH,AO,OW,UH/.

The average percentage of agreement between identifications collected from each subject response set paired with all other subject response sets for all 1725 tokens across all confidence ratings are shown in Table 2.2. The average percentage of agreement across all subjects was 63.8% and drops to 62.9% when the agreements of subjects with their own replications are not included. These averages and their standard deviations for agreement with others are plotted in Figure 2-6 for all subject response sets. The average agreement across subjects for the first response sets was 61.6% and for the second response sets increased to 64.5%.

To test differences in agreement, the statistic kappa (κ) (Cohen, 1960) was utilized. This statistic provides a coefficient of agreement between two raters for nominal scales and includes consideration of chance agreements. The statistic assumes that all disagreements may be considered equally serious. A coefficient of 1.00 reflects perfect agreement between raters, while a coefficient of zero reflects total independence between the two raters. When N is large (i.e. > 100),the sampling distribution of κ approximates normality. Thus differences between two values of κ may be tested for significance by evaluating the normal curve

Table 2.1: Frequency of ID responses by subject response set.

Set	IY	IH	EH	EY	AE	AA	AH	AO	OW	UH	UW	ER
1M1	53	63	63	22	88	174	130	170	146	269	397	150
1M2	69	63	89	41	100	167	129	170	151	256	394	96
3M1	135	187	73	9	68	166	176	138	130	223	374	46
3M2	136	164	69	14	56	132	161	144	153	232	385	79
4M1	269	122	77	17	102	142	104	152	171	108	245	216
4M2	264	143	95	18	102	118	113	146	183	128	271	144
5M1	239	106	52	149	158	186	120	122	86	120	306	81
5M2	269	125	47	119	141	194	116	117	129	87	314	67
1F1	82	25	40	37	83	123	130	256	120	225	513	91
1F2	178	116	65	22	77	141	120	192	137	196	398	83
2F1	217	73	86	64	94	133	127	144	171	156	377	83
2F2	289	59	106	83	81	125	116	135	199	160	331	41
3F1	95	66	99	16	70	113	164	143	161	312	375	111
3F2	90	61	102	21	58	92	155	175	166	253	454	98
5F1	174	14	118	89	85	133	197	160	180	56	448	71
5F2	203	20	117	79	108	152	173	132	198	85	413	45
$\overline{x}$	173	88	81	50	92	143	139	156	155	179	375	94

deviate.

Kappas and standard errors of κ were calculated for each subject's first response set paired with all other first response sets and similarly for subjects' second response sets. The average value of κ for the first response sets was .570 and for the second response sets was .602. The difference between these values was not found to be significant ($z = 1.78$), suggesting that while subjects' agreements improved in the second set, they were not significantly different from the first set agreement performance.

To consider test-retest reliability, the percentage of agreement between each subject's first and second response set was first calculated. The average of these agreements was

75.31%, approximately 12% greater than the overall average of agreements between each subject and all others. Kappas and standard errors of κ were calculated for each subject's response set paired with all other response sets except the subject's replication. The average across all subjects was $\kappa = .585$. The same statistics were calculated for each subject and their replication, yielding an average $\kappa = .721$. The difference between these two values of κ was tested for significance by evaluating the normal curve deviate. The difference was highly significant ($z = 7.76, p < .001$). This result suggests that test-retest reliability within subjects is quite high when compared to average agreement between subjects.

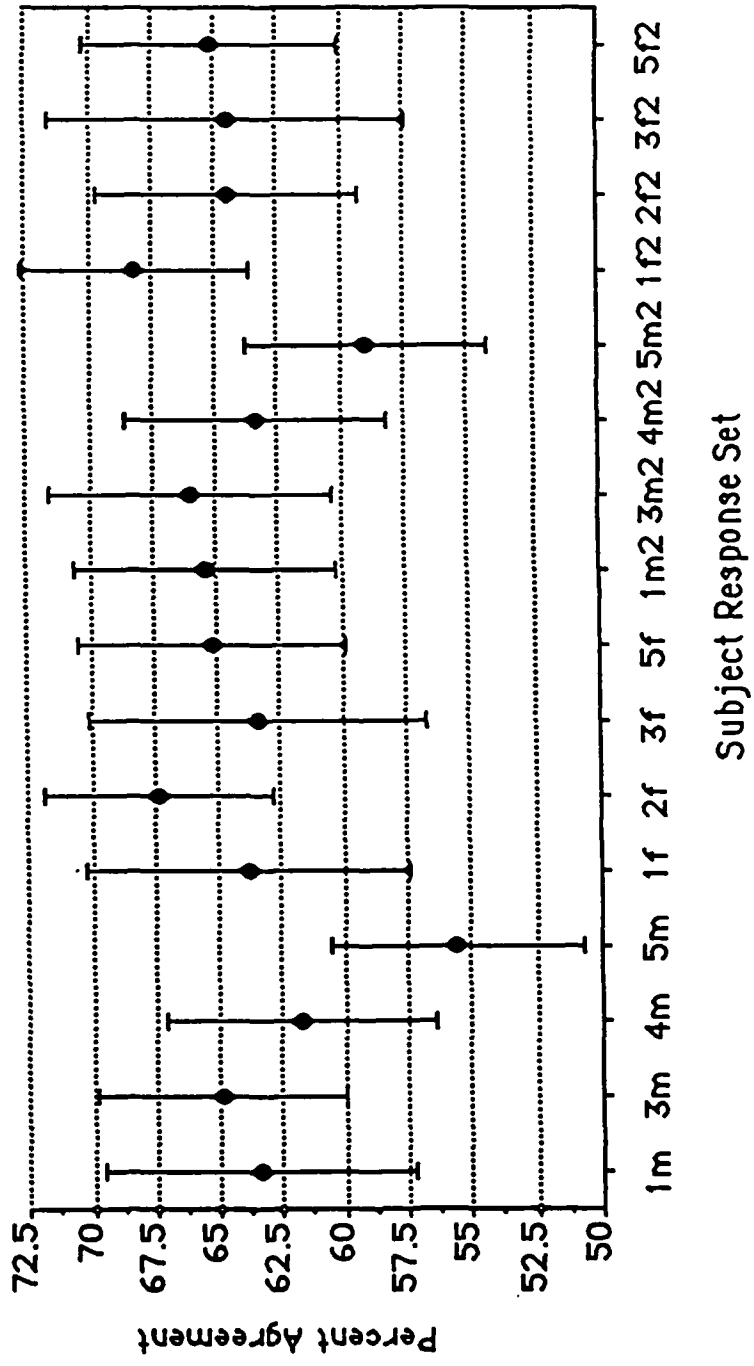
2.3.3 Ratings

The frequencies of rating responses by rating category are shown in Table 2.3 for each subject response set. It is readily apparent that the distribution of rating responses is negatively skewed with the highest frequency of ratings occurring in the fourth category representing a confidence level of greater than mediocre but less than very sure. It does appear that subjects were generally more sure of their responses than less sure. The variations in the frequency of rating responses within subjects is smaller than the variations between subjects. This suggests that each subject establishes his/her own criteria for ratings judgements, and maintains the same criteria run to run. This notion is consistent with data reported in a study investigating the use of confidence ratings for identifications of spondee words presented in noise by F.R. Clarke (1960). As was previously reported by Pollack and Decker (1958), Clarke noted that listeners tended to underestimate their ratings of high

Table 2.2: Percentages of identification agreement by subject response set.

Run	1M	1M2	3M	3M2	4M	4M2	5M	5M2	1F	1F2	2F	2F2	3F	3F2	5F	5F2
1M	—															
1M2	75.4	—														
3M	65.1	67.7	—													
3M2	66.1	67.9	76.2	—												
4M	58.2	59.4	60.6	61.2	—											
4M2	59.0	60.9	64.2	65.9	76.4	—										
5M	52.2	56.0	55.8	54.2	55.8	55.8	—									
5M2	55.4	58.7	59.3	57.9	58.9	59.8	70.4	—								
1F	70.4	70.7	64.3	65.7	57.0	58.5	51.9	54.0	—							
1F2	70.1	71.5	72.0	73.8	65.0	67.0	57.5	60.8	71.9	—						
2F	63.8	66.3	67.2	69.9	67.7	69.0	57.9	61.7	65.2	70.5	—					
2F2	58.2	63.1	64.5	66.6	66.0	67.0	56.5	60.8	59.4	67.8	76.8	—				
3F	65.8	66.1	63.1	67.0	59.5	61.8	50.0	51.7	66.2	66.4	65.2	60.2	—			
3F2	66.5	68.0	65.6	68.9	59.1	61.0	48.9	51.5	71.8	68.6	66. 6	62.4	76.2	—		
5F	62.6	64.8	63.5	64.5	61.4	61.2	55.5	61.2	65.6	67.6	70.9	68.8	63.9	66.8	—	
5F2	61.6	64.7	64.0	63.9	61.7	63.0	57.3	63.4	64.1	67.0	71. 0	68.8	62.7	65.0	79.2	—
$\bar{x}$	63.4	65.4	64.9	66.0	61.8	63.4	55.6	59.0	63.8	67.8	67. 3	64.4	63.0	64.4	65.2	65.2

Figure 2-6: Mean agreement between each subject response set and all other response sets on identification of the 1725 synthetic vowels. Error bars indicate ±1 standard deviation.



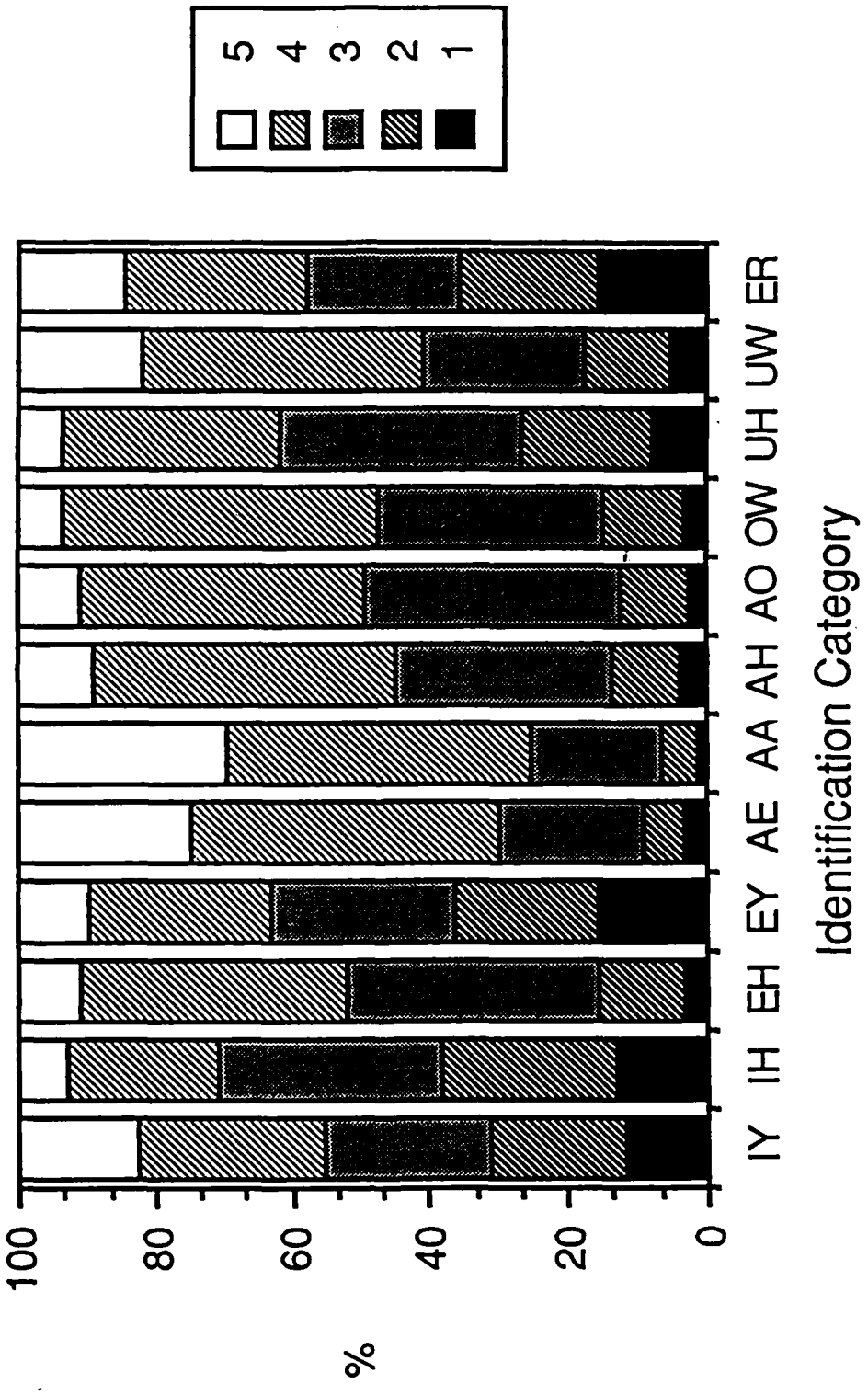
intelligibility items (and overestimate their ratings of low intelligibility items). A pattern of this nature could account for the high frequency of ratings in category 4 of the present data if subjects are, in fact, underestimating their confidences at identifying tokens of high salience.

Table 2.3: Frequency of rating responses by subject response set.

Set	1	2	3	4	5
1M	63	207	444	695	316
1M2	47	164	370	647	497
3M	87	369	636	548	85
3M2	73	349	680	548	75
4M	214	257	371	692	191
4M2	193	201	407	735	189
5M	236	418	373	360	338
5M2	171	325	408	380	441
1F	34	132	488	618	453
1F2	9	130	424	658	504
2F	54	183	441	725	322
2F2	57	193	429	843	203
3F	220	273	679	549	4
3F2	83	213	509	909	11
5F	120	197	552	749	107
5F2	90	156	469	757	253
Total	1751	3767	7680	10413	3989
$\bar{x}$	109.4	235.4	480.0	650.8	249.3

Figure 2-7 shows the frequencies of rating responses for each identification category expressed as percentages of all ratings for each category. The figure indicates that the majority of tokens across all categories received a rating of 3 or 4, as was noted above,

Figure 2-7: Percentage of subjects' confidence rating responses by individual identification category.



but also, that tokens classified as the "point" vowels /IY,AE,AA,UW/ and the retroflex /ER / received greater than 15% of their ratings in the "very sure" (5) rating category, suggesting that a greater percentage of tokens classified as these vowels were very salient as opposed to tokens classified as more centralized vowels. Additionally, tokens classified as /IY,IH,EY,UH,ER/ received greater than 25% of their ratings in rating categories 1 and 2. These higher percentages of low confidence ratings most probably reflect observations noted in Section 2.3.1, that is, for the /IY/ category, a number of tokens probably most representative of a rounded front vowel /ü/ not used in American English may have been classified as /IY/. Likewise, uncertainty in classifying /EY/ may reflect that many subjects expressed difficulty in differentiating /EY/ from neighboring vowels. Classifying tokens in the /ER/ region may present a higher percentage of uncertainty in that tokens synthesized near the /ER/ boundary take on a high degree of retroflexion and are very "r-colored", creating ambiguities. The lax vowel categories, in particular /IH/, have often appeared less stable in terms of perceptual identification (See footnote 2, Section 2.2.1) than are other vowel categories of English, suggesting that a higher percentage of uncertainty may be associated with them.

The percentage agreement on confidence ratings between each subject response set and every other response set is shown in Table 2.4. The overall average agreement is 30.2%, considerably less than their agreement on token identifications. This overall agreement drops to 29.6% when the agreements of subjects with their own replications are not included. The average value of κ for these agreements is quite small, .049, suggesting that a large portion of agreements may now be random chance. Subjects' average percentage of agreement was uniform across the first (29.2%) and second (30.0%) response sets. Once again, the agreement within subjects (38.6%) was significantly higher than the overall average based on a comparison of the average κ statistic for each condition ($z = 4.38, p < .001$).

In summary, the data from confidence ratings suggests that subjects establish individual criteria for basing their confidence judgements, and then use these criteria consistently. These individual differences are reflected in the lower agreements between subjects found for confidence ratings as compared to identification agreements. The majority of ratings fall in category 4, reflecting a confidence judgement which is greater than mediocre but

Table 2.4: Percentages of agreement on confidence ratings by subject response set.

Run	1M	1M2	3M	3M2	4M	4M2	5M	5M2	1F	1F2	2F	2F2	3F	3F2	5F	5F2
1M	—															
1M2	38.2	—														
3M	32.3	27.5	—													
3M2	32.5	25.3	44.0	—												
4M	31.4	27.0	30.1	28.8	—											
4M2	31.7	27.8	30.3	31.2	36.1	—										
5M	28.3	26.7	29.0	25.8	25.2	25.2	—									
5M2	27.7	26.4	25.4	26.1	25.0	23.6	36.3	—								
1F	32.2	34.7	27.1	27.3	26.5	28.1	23.2	21.9	—							
1F2	35.9	36.8	28.3	27.0	28.8	31.1	25.7	26.6	38.1	—						
2F	36.4	37.8	31.9	31.2	32.1	31.5	25.0	26.1	35.5	34.7	—					
2F2	35.4	33.9	35.0	32.5	33.2	35.2	26.2	26.4	33.0	35.5	43.2	—				
3F	28.1	23.5	32.1	33.7	28.6	29.8	21.9	22.4	25.2	23.9	27.4	29.8	—			
3F2	33.0	28.4	35.4	35.7	32.3	33.9	21.8	20.9	30.7	29.5	32.9	35.8	37.4	—		
5F	32.1	29.4	30.1	31.7	29.4	34.5	22.3	20.2	29.6	28.3	32.8	35.1	30.6	36.6	—	
5F2	32.3	28.9	28.8	29.2	29.4	31.7	20.9	22.9	32.5	30.6	34.5	35.7	29.2	33.6	35.2	—
$\bar{x}$	33.8	30.1	31.2	30.8	29.6	30.8	25.6	25.2	29.7	30.7	32. 9	33.7	28.2	31.9	30.5	30.4

less than very sure. Results of past studies, however, suggest that subjects may tend to underestimate their confidence when they are correct. The identifications assigned to vowel categories representing the extreme points of articulation, (/IY,AE,AA,UW,ER/) tended to receive higher confidence ratings than did vowel tokens in categories representing less extreme articulatory points.

2.3.4 Synthetic speech-based (*SSB*) target zones

Consideration is next given for whether target zones similar to those illustrated in Figures 1-8 and 1-9 from Chapter 1 can be constructed in the *APS* based on the identifications for the 1725 synthetic tokens. A single vowel category classification for each stimulus token is required for constructing such zones. The vowel category representing the plurality of identification responses for each token was determined for this purpose. That is, for each point in the *APS* representing one of the tokens, the vowel identification category receiving the greatest number of the 16 (8 subjects x 2 replications) possible responses for that token was taken as the category label for that point.

Boundary lines were drawn enclosing points of like plurality identification in each z' plane. Each point was considered to represent the space occupied by a 0.05 log unit cube with the point at its center. Adjacent cubes of unlike categories were separated by boundary lines drawn equidistantly between the two cubes such that each line was .024 log units away from its associated cube center in the $x'y'$-plane. This technique successfully enclosed 1674 of the token points into adjacent, non-overlapping zones along each of the five z' planes and are shown in Figures 2-8a-g.

Of the 51 token points remaining, 49 represented ties between the greatest number of subjects' responses for two or more categories such that no clear plurality could be established. Of the tied token points, 3 represented ties between three identification categories and the remaining 46, ties between two categories. All but one of the tied token points fell along boundary areas between the adjacent zones which were representative of the identification categories of the ties. These points were included in the zone construction by drawing boundary lines diagonally to either side of the point in the $x'y'$-plane or otherwise dividing the point's cube such that all tied response categories associated with that cube

Figure 2-8: (a) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.80$ plane.

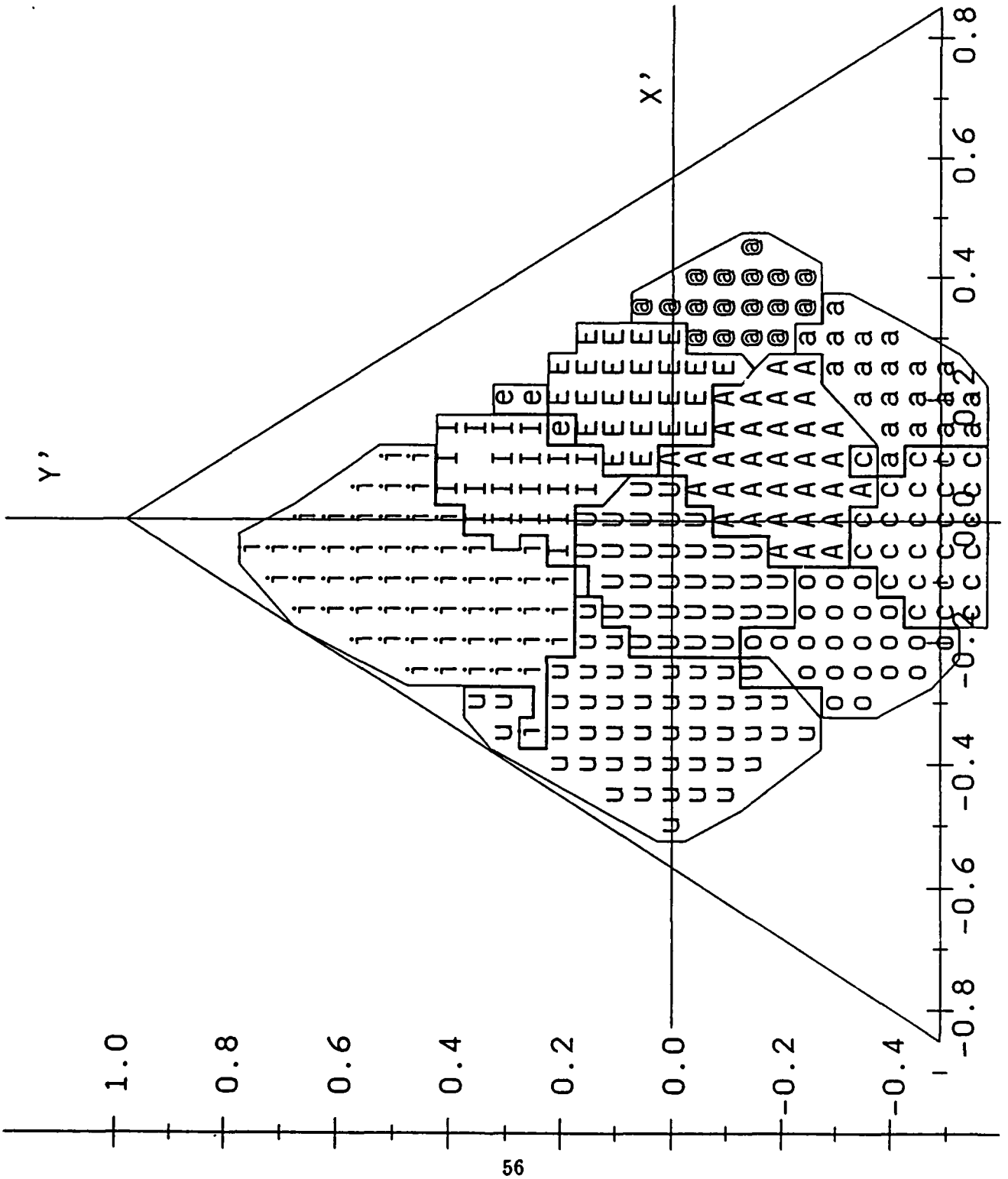


Figure 2-8: (b) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.75$ plane.

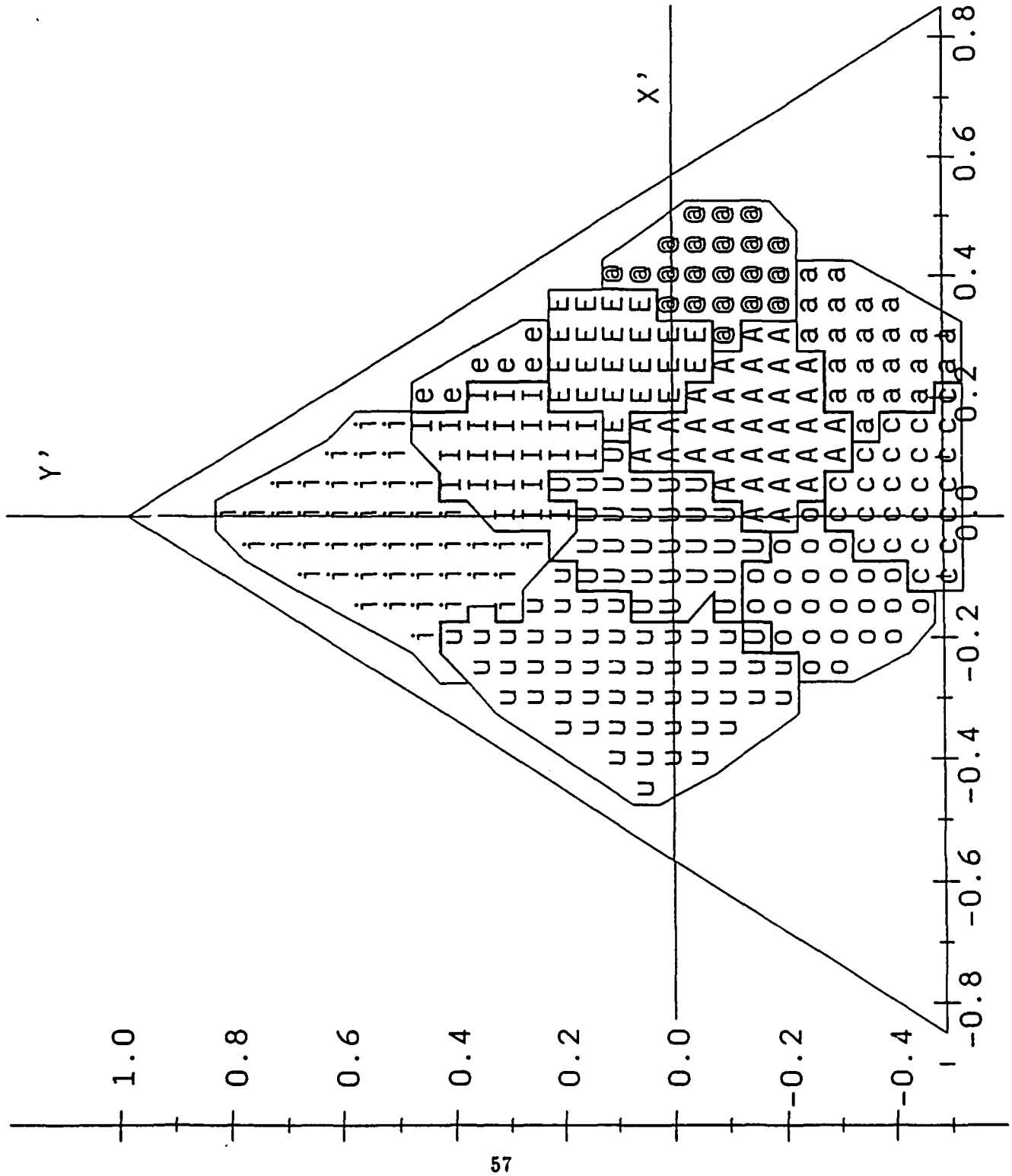


Figure 2-8: (c) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.70$ plane.

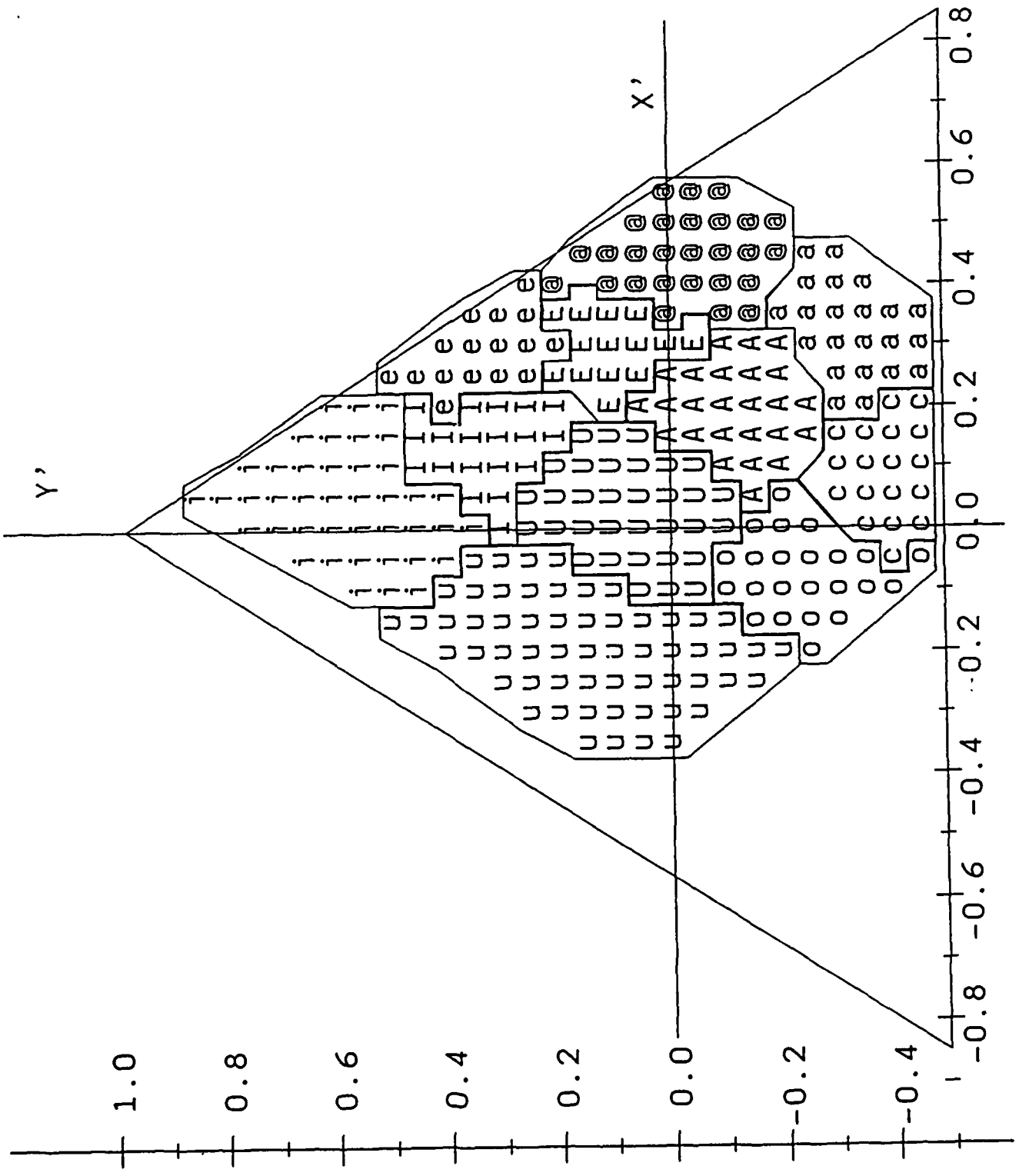


Figure 2-8: (d) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.65$ plane.

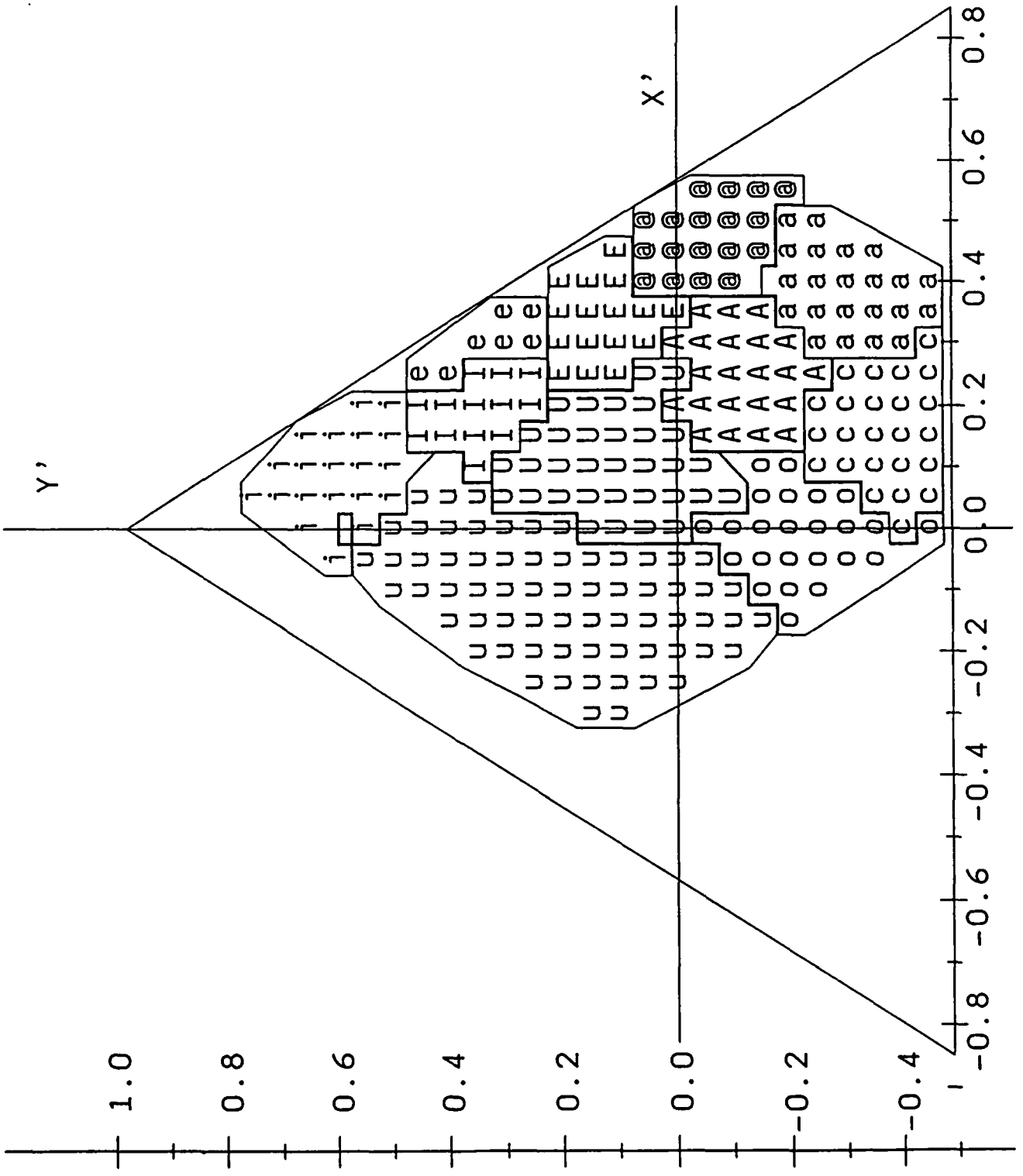


Figure 2-8: (e) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.60$ plane.

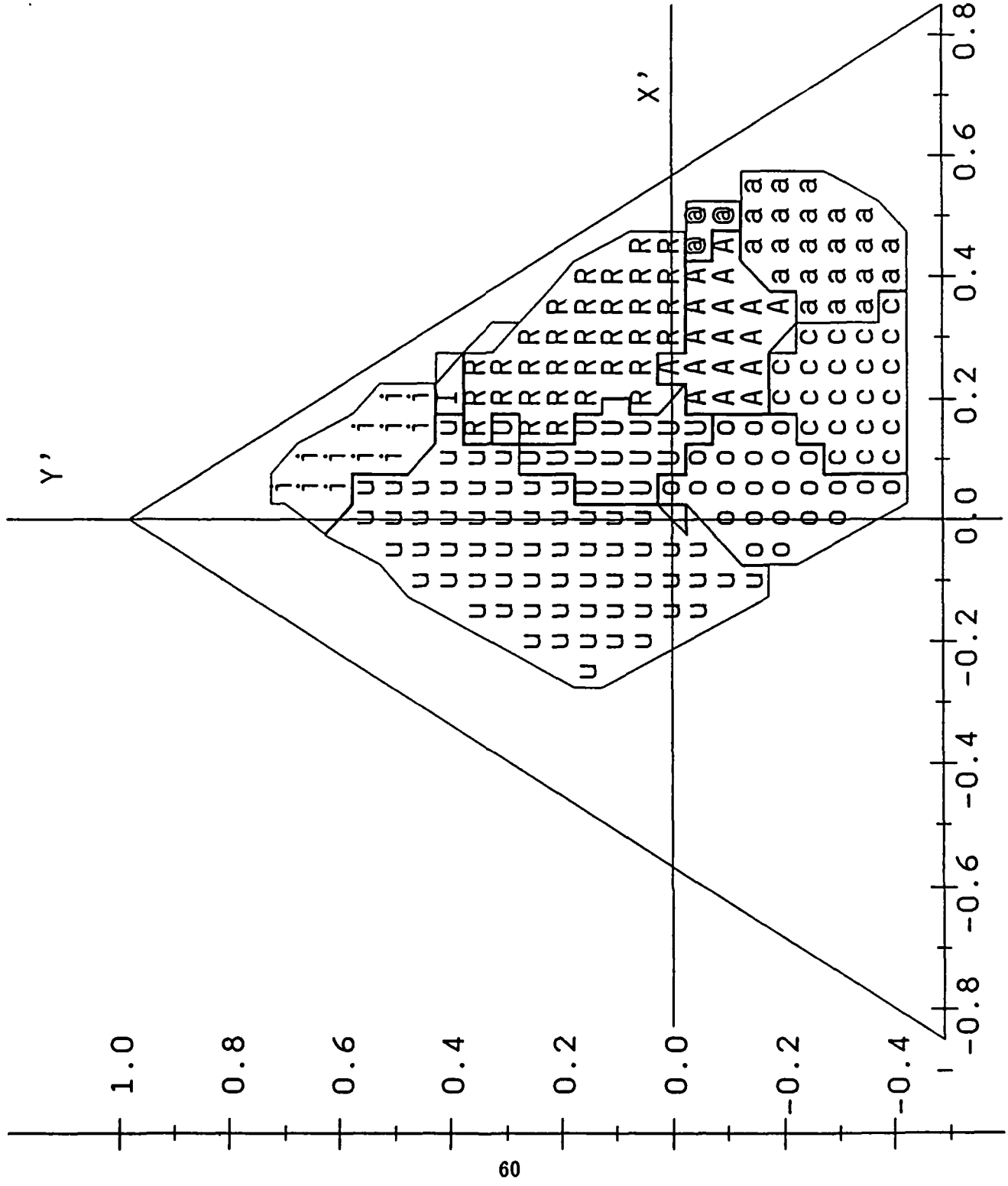


Figure 2-8: (f) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.55$ plane.

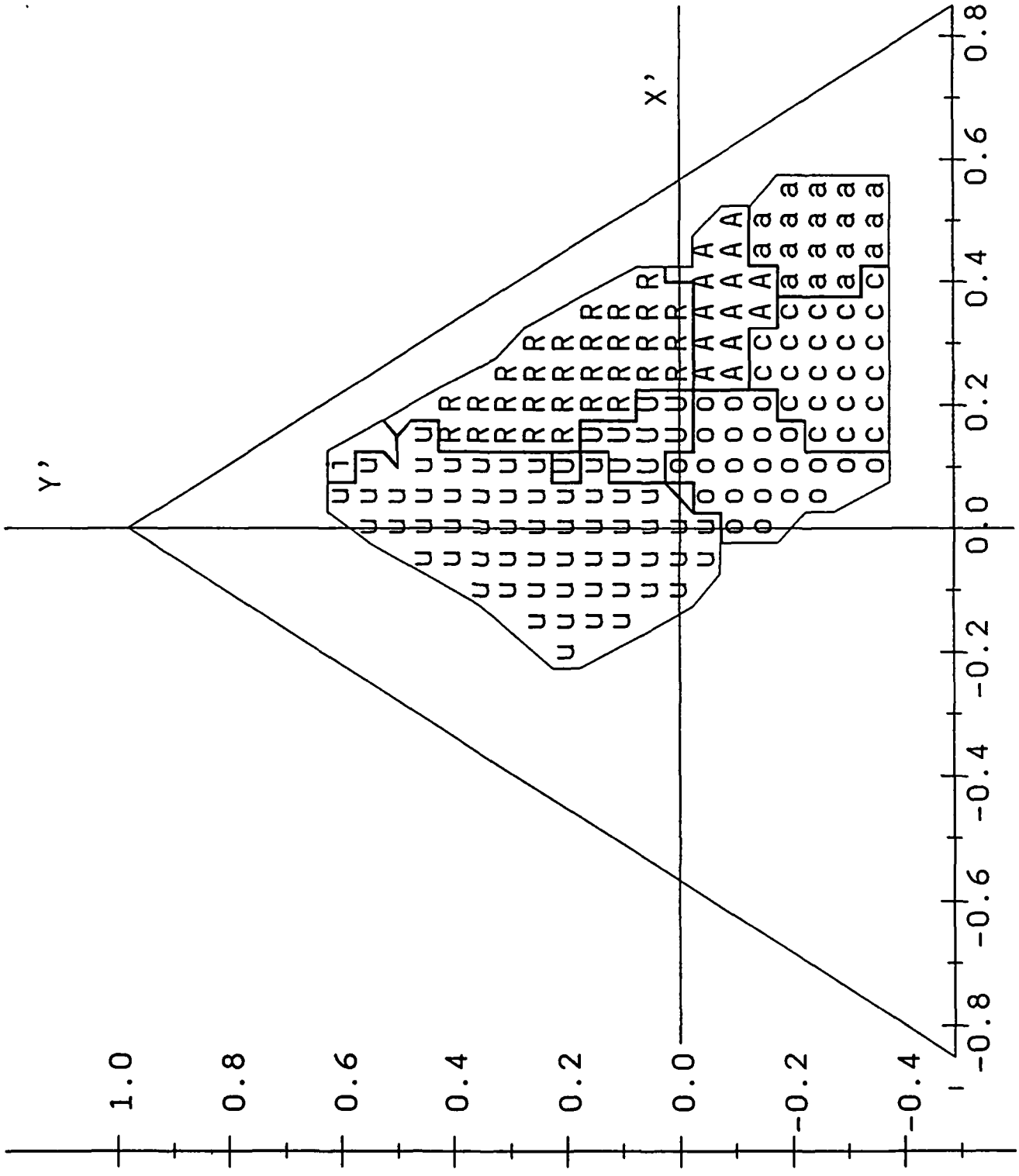
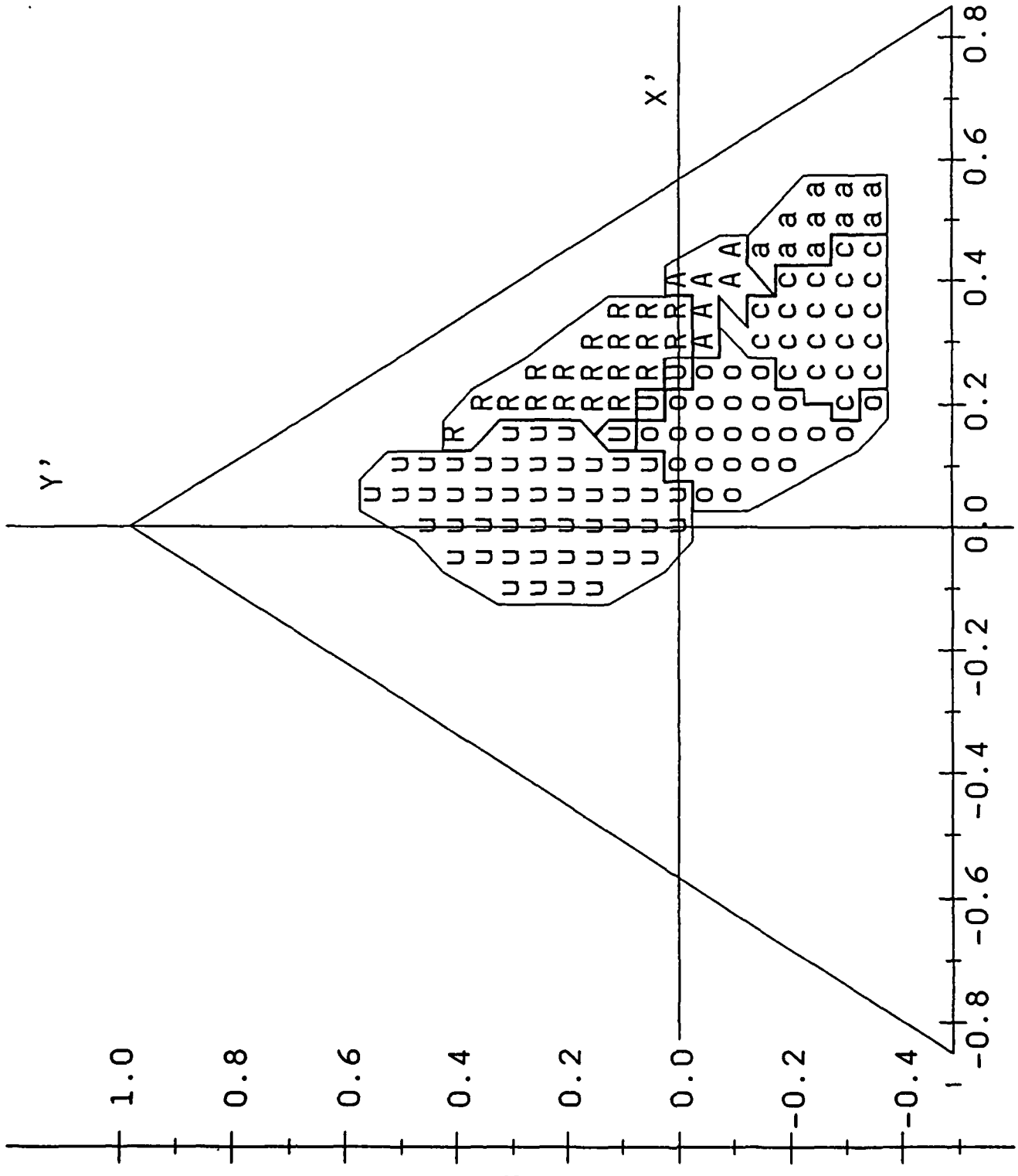


Figure 2-8: (g) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.50$ plane.



were approximately equally represented. One tied token (See $x' = 0.1$, $y' = 0.4$, $z' = 0.65$, Fig. 2-8d), could not be represented with a boundary line without causing overlap, so its identification was assigned to one of the two tying categories. It should be noted however that while the cubes associated with the tied points are divided between several identification categories, the exact point locations in the *APS* of these tied tokens do not fall into any zone, but rather, fall between zones.

The two remaining tokens (see Fig. 2-8a) could not be associated with a zone based on the plurality responses without creating overlap in the zones. For these cases, the response category receiving the second highest number of responses for each of these tokens corresponded with an associated zone such that no overlap would be created and was thus chosen as the correct identification for purposes of constructing the zones.

In summary, it is clear that zones for vowel cateogories can be constructed when the identifications for the plurality of subjects are utilized. These zones are adjacent and non-overlapping, and successfully account for over 99% of the synthetic tokens when they are represented as points in the *APS*. In future sections, data analyses pertaining to plurality rates and plurality identifications often will not include the 51 tied or overlapping points discussed above in the data set. These points will be referred to at those times as the "rejected" points.

2.3.5 Qualitative analysis of synthetic speech-based target zones

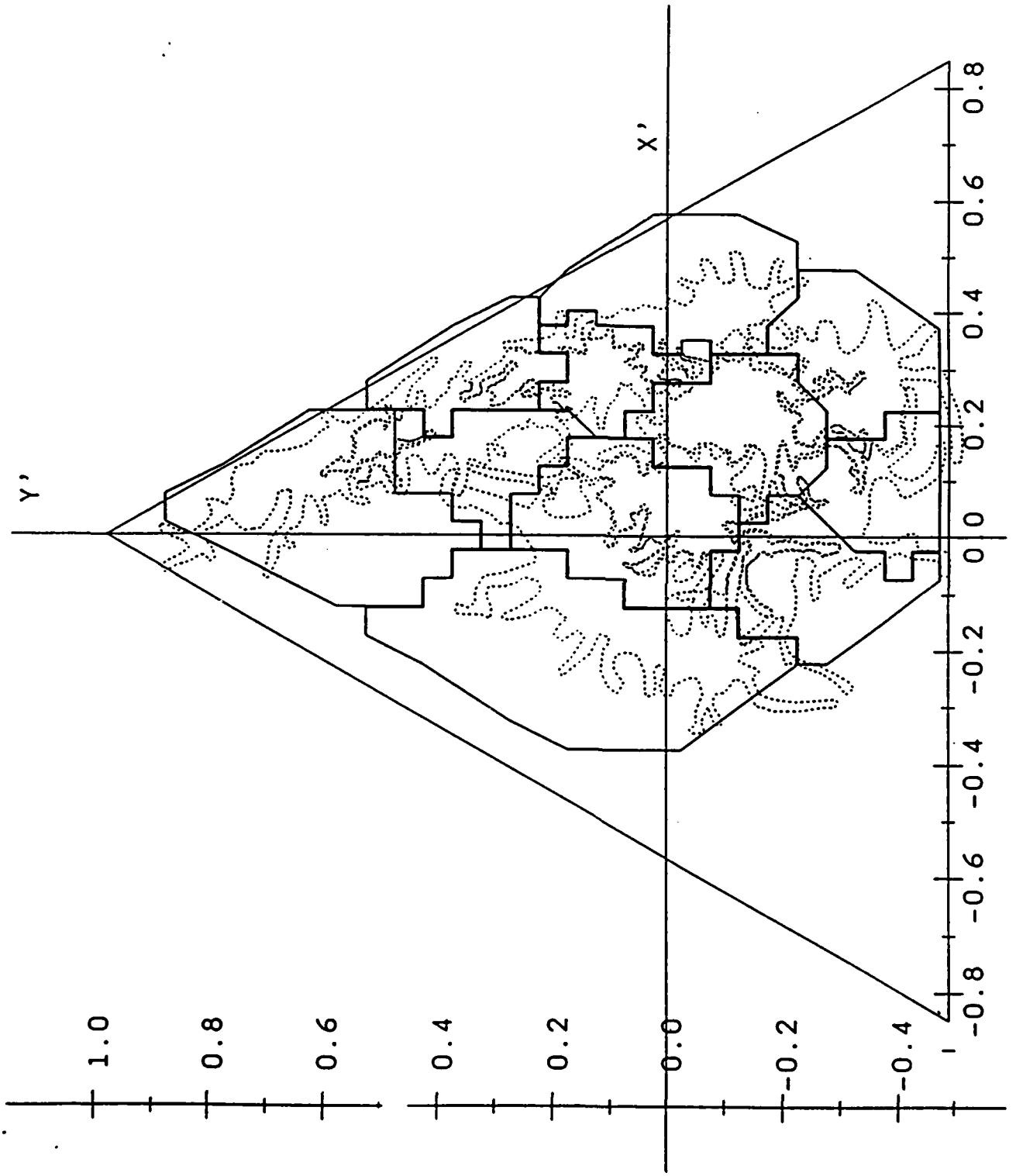
Upon examination of Figures 2-8a-g, it is first apparent that the overall area occupied by the zones is not uniform through z', but changes in shape, shifts in location, and gradually becomes smaller with decreasing values of z'. These changes in overall area reflect the second set of criteria (See Section 2.2.1) utilized in limiting the stimuli to acceptable formant combinations based on production restraints found in natural speech , and therefore, reflect not only these natural constraints, but also define the space available for synthesis. Notice also that the zone boundaries are not stable in terms of the z' dimension, but rather, change with each plane through z'. Additionally, the zones themselves appear to shift with decreasing changes in z' in an upwardly direction and to the right. This apparent shift suggests that like identifications are generally following lines of constant $F1$ and $F2$ back

through z' as is illustrated and discussed in Appendix A.

Figure 2-9 shows the *SSB* zones for the central primary ($z' = 0.70$, from Figure 2-8c) plane overlayed on the most recently estimated zones based on natural-speech productions from Figure 1-9, Chapter 1. The locations of the new *SSB* target zones relative to one another is approximately the same as the locations of the target zones based on natural speech. The locations of the vowel categories /EY/ and /OW/ not used in the natural speech-based (*NSB*) target zones fall predominantly into areas of unclaimed space relative to the *NSB* zones, that is, /EY/ lies to the right of /IH/ and between /IY/ and /EH/, /OW/ lies between /AO/ and /UW/ and is bordered by /AH/ and /UH/. These locations are in agreement with traditional drawings of vowels according to articulatory tongue position. The zone for the retroflex /ER/ lies behind the non-retroflex vowels on the z' axis and ranges from $z'= 0.50$ to 0.60. At $z'= 0.65$, the zones for /IH/, /EH/, and /UH/ emerge in the same $x'y'$-space as /ER/, suggesting that the lower z' values are associated with shifting the percept to /ER/. Lower z' values in *APS* translate to lower $F3$ values which have often been associated with perception of retroflexion. The fact that overlap is found between the /ER/ and the /IH/, /EH/, and /UH/ zones when z' is not considered is in general agreement with plots of $F1$ by $F2$ for vowels from Peterson and Barney (1952) which also showed overlap for these categories.

In summarizing this section, we find that the *SSB* zones for vowels in the *APS* change in shape and location as z' changes. These changes reflect the area available for synthesis and the fact that like values of $F1$ and $F2$ are generally associated with the same vowel category, except in the case of perceived retroflexion. The *SSB* zones in the primary planes are in general agreement with the *NSB* zones, with the new *SSB* zones for /EY/ and /OW/ predominantly occupying what was considered unclaimed space with the *NSB* zones. The zone for /ER/ falls in a distinct area behind the zones for /IH/, /EH/, and /UH/, supporting the generally accepted notion that the percept of /ER/ is primarily mediated by a lowered $F3$.

Figure 2-9: Synthetic-speech-based target zones (solid lines) from Figure 2-8 and natural-speech-based target zones (dashed lines) from Figure 1-3 for the $z' = .70$ plane.



2.3.6 Plurality agreements on identifications

The number of subject responses constituting a plurality, the plurality frequency, varied from 5 to 16 out of the 16 possible responses per token. The number of tokens and the percentage of the total number of tokens by plurality frequency are shown in Table 2.5 for all z' planes and for the primary planes only. For example, the row corresponding

Table 2.5: Agreement on identification responses by plurality frequency.

Plurality Frequency	All Planes	%	Primary Planes	%
16/16	320	18.6	170	19.3
15/16	190	11.0	103	11.7
14/16	171	9.9	94	10.7
13/16	148	8.6	75	8.5
12/16	167	9.7	88	10.0
11/16	137	7.9	63	7.1
10/16	145	8.4	77	8.7
9/16	154	8.9	74	8.4
8/16	128	7.4	68	7.7
7/16	71	4.1	34	3.9
6/16	34	2.0	14	1.6
5/16	11	0.6	3	0.3
Total	1676	97.2	863	97.8

to a plurality frequency of 16 indicates the number and percentage of tokens where all subjects' identifications agreed. We find here that all identifications agreed on 320 tokens across all z' planes and 170 tokens in the *primary* planes. Note that, in general, plurality agreements tend to be quite high with almost 20% of the tokens unamimously agreed upon and almost 60% agreed upon by 75% (plurality frequencies of 12 and greater) or more of the identifications. The percentage of tokens for each plurality frequency does not appear to change significantly when limited to tokens only in the primary planes. This suggests that patterns of plurality agreement are relatively unaffected by values of $F3$ outside the normal $F3$ ranges found in natural speech for non-retroflex vowels.

The plurality frequencies for each token (excluding rejected points) are shown in Figures

2-10a-g for each of the seven z' planes, along with the *SSB* zones based on the plurality identifications. Note that the higher plurality frequencies tend to lie to the interior of each zone and smaller frequencies nearer the boundaries. Additionally, comparison of any given target zone across z' planes shows that some z' planes contain generally higher plurality frequencies than others. If plurality frequencies are considered to be representative of the saliency of tokens, with higher values representing greater saliency and lower values representing less saliency, we find that, not only do the target zones vary in their salience relative to one another, but also that a salience gradient may be applied in all three dimensions of each target zone. Such a gradient could be used to estimate the relative likelihood of associating a given point in *APS* with a particular vowel category. Thus a point falling in a region of high salience could be assigned to the vowel category representing that region with relatively high certainty and a point falling in a region of low salience could receive a low certainty identification or multiple, ambiguous identifications.

2.3.7 Confidence ratings

Can confidence ratings be used in addition to plurality frequencies to establish a saliency gradient? Clarke (1960) suggests that confidence ratings may increase the amount of information transmitted an additional 16.5% over identification responses alone. One approach toward answering this question is to determine whether or not confidence ratings are correlated with plurality frequencies. Figures 2-11a-g show the sum of all 16 confidence ratings for each token plotted in the *APS* along with the *SSB* target zones for all seven z' planes. These values have a possible range from 16, were all subjects to assign a token a 1 rating, to 80, were all subjects to assign a token a 5 rating. As with the identification pluralities, we find that, for a given z' plane, larger values generally lie more interior to the target zones and smaller values nearer the boundaries and that values also vary as a group for a single target zone across z' planes. A non-parametric statistical procedure, the Spearman rank-order correlation, was used to measure the degree of association between confidence rating sums and identification plurality numbers. A moderate correlation ($R^2 = .508$) after correcting for ties was found. Figure 2-12 shows the means and standard deviations for the confidence rating sums grouped by plurality number. Thus, while confidence ratings

Figure 2-10: (a) Plurality frequencies (See text) for all tokens in the $z' = 0.80$ plane.

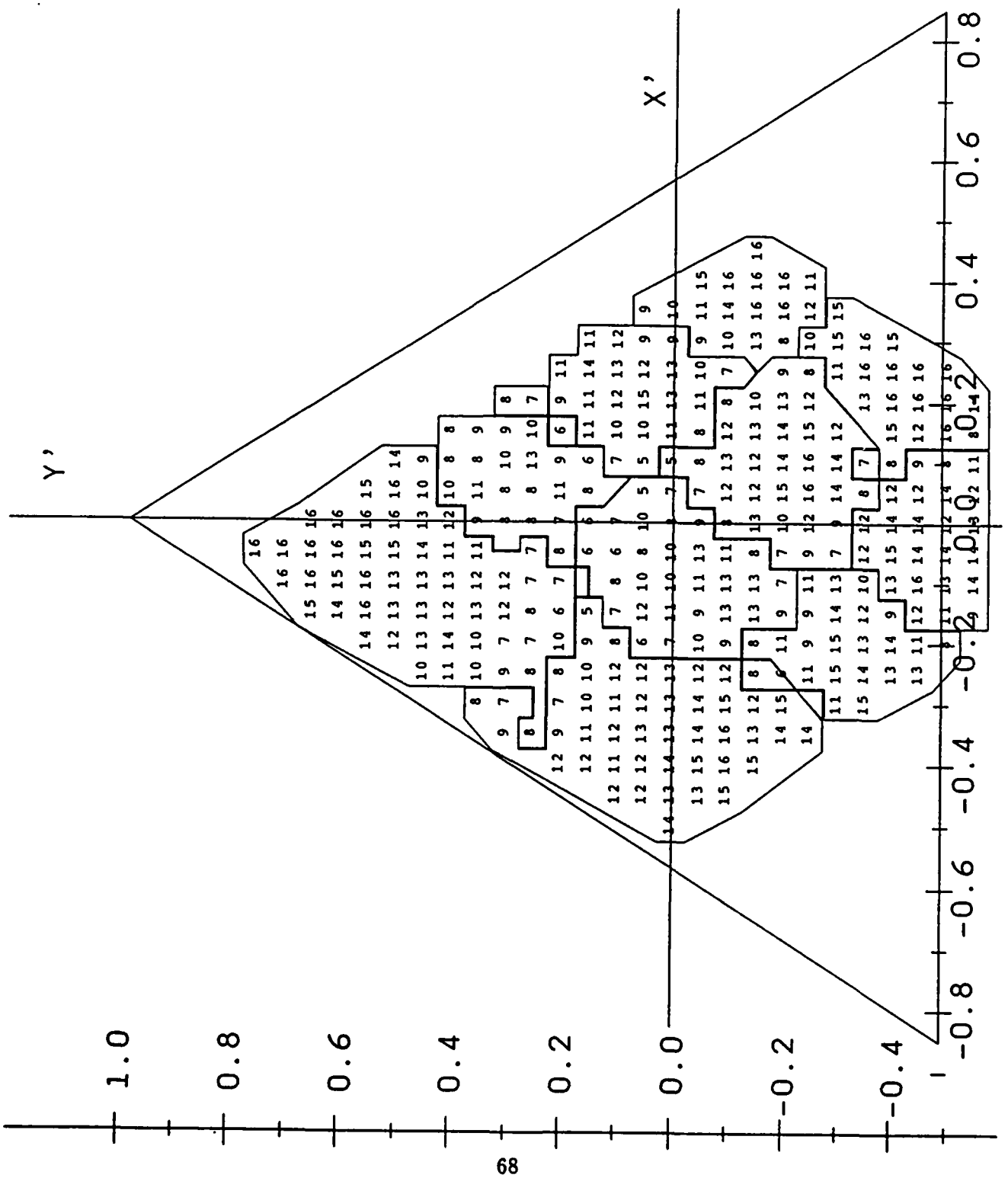


Figure 2-10: (b) Plurality frequencies (See text) for all tokens in the $z' = 0.75$ plane.

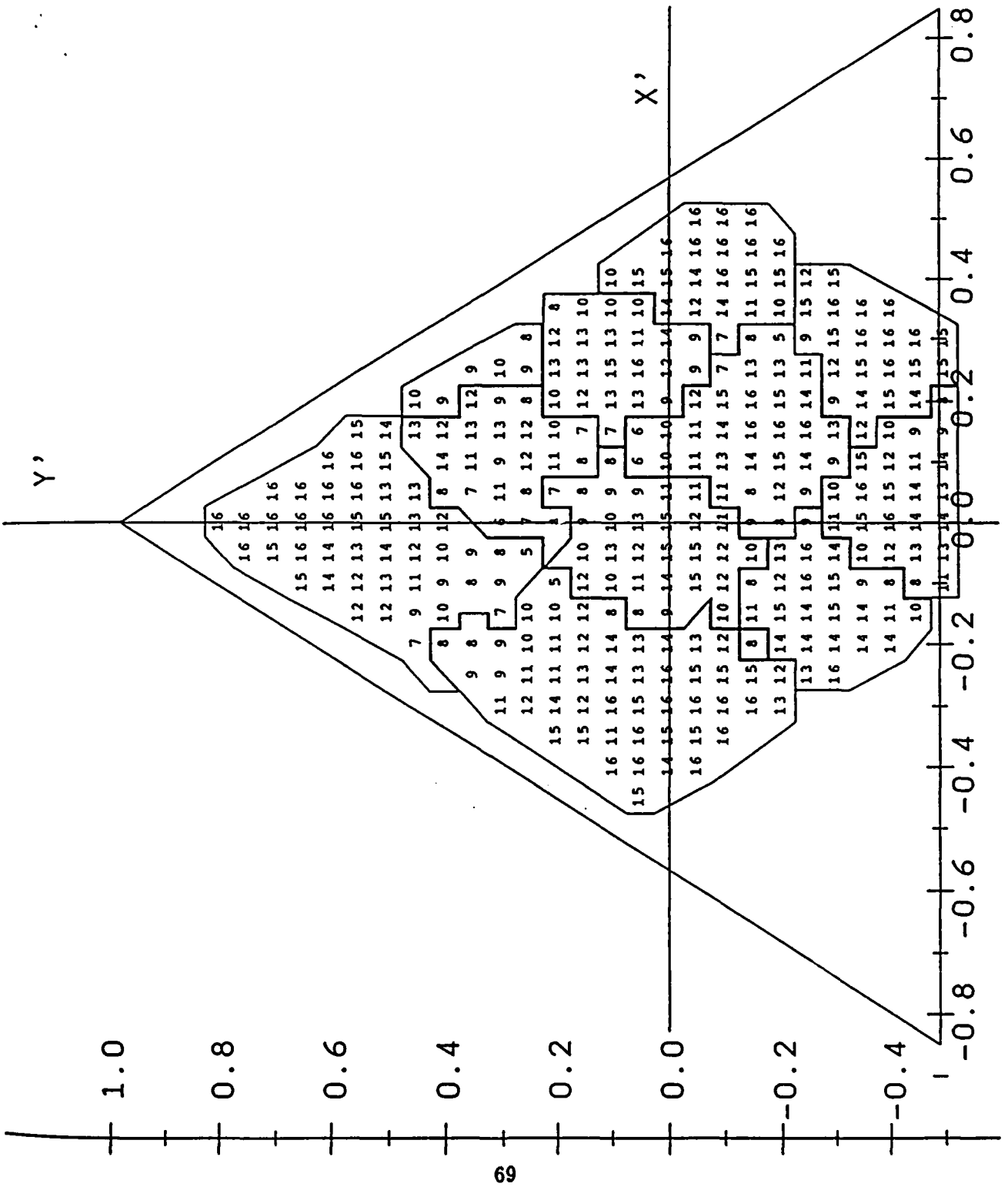


Figure 2-10: (c) Plurality frequencies (See text) for all tokens in the $z' = 0.70$ plane.

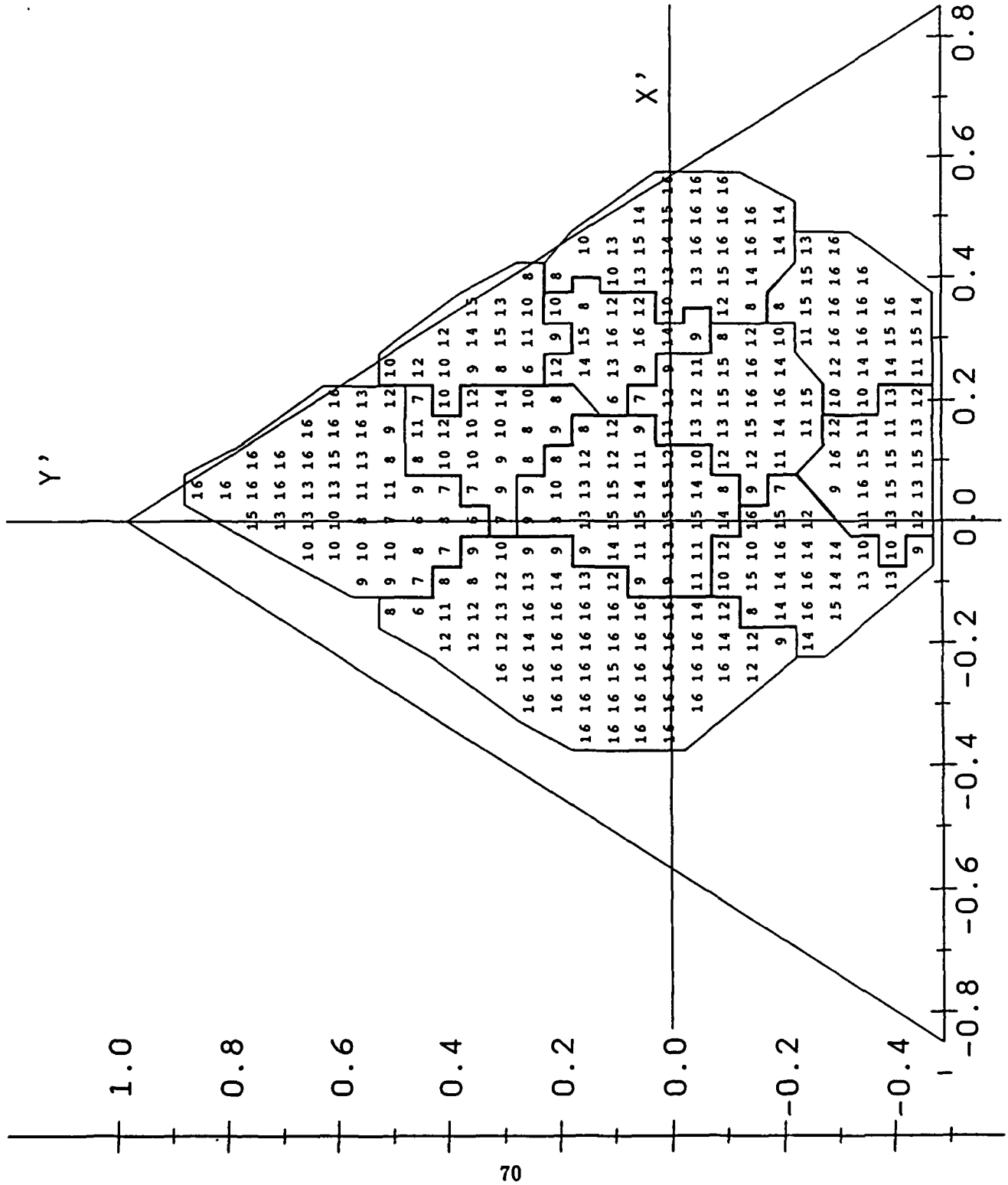


Figure 2-10: (d) Plurality frequencies (See text) for all tokens in the $z' = 0.65$ plane.

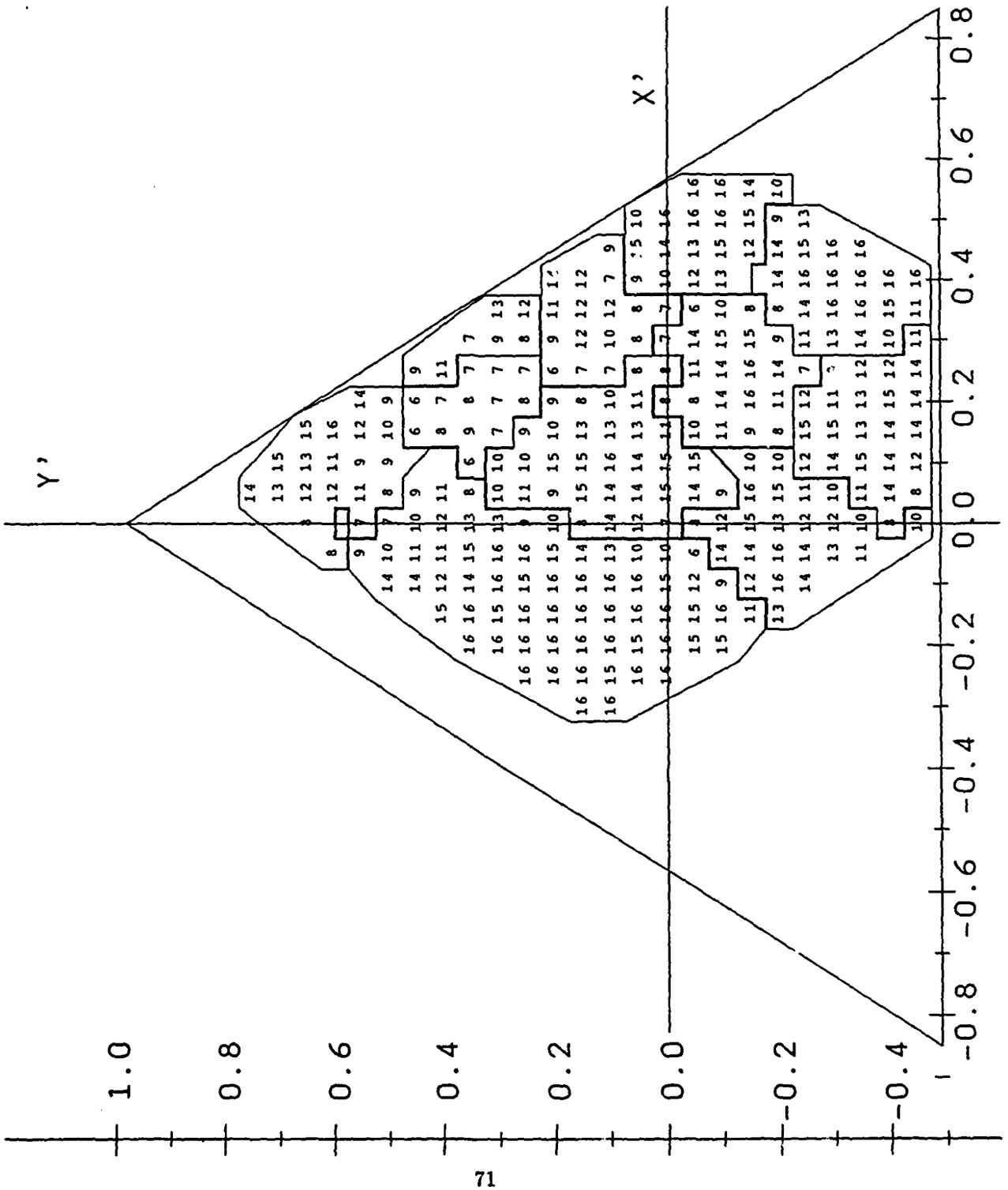


Figure 2-10: (e) Plurality frequencies (See text) for all tokens in the $z' = 0.60$ plane.

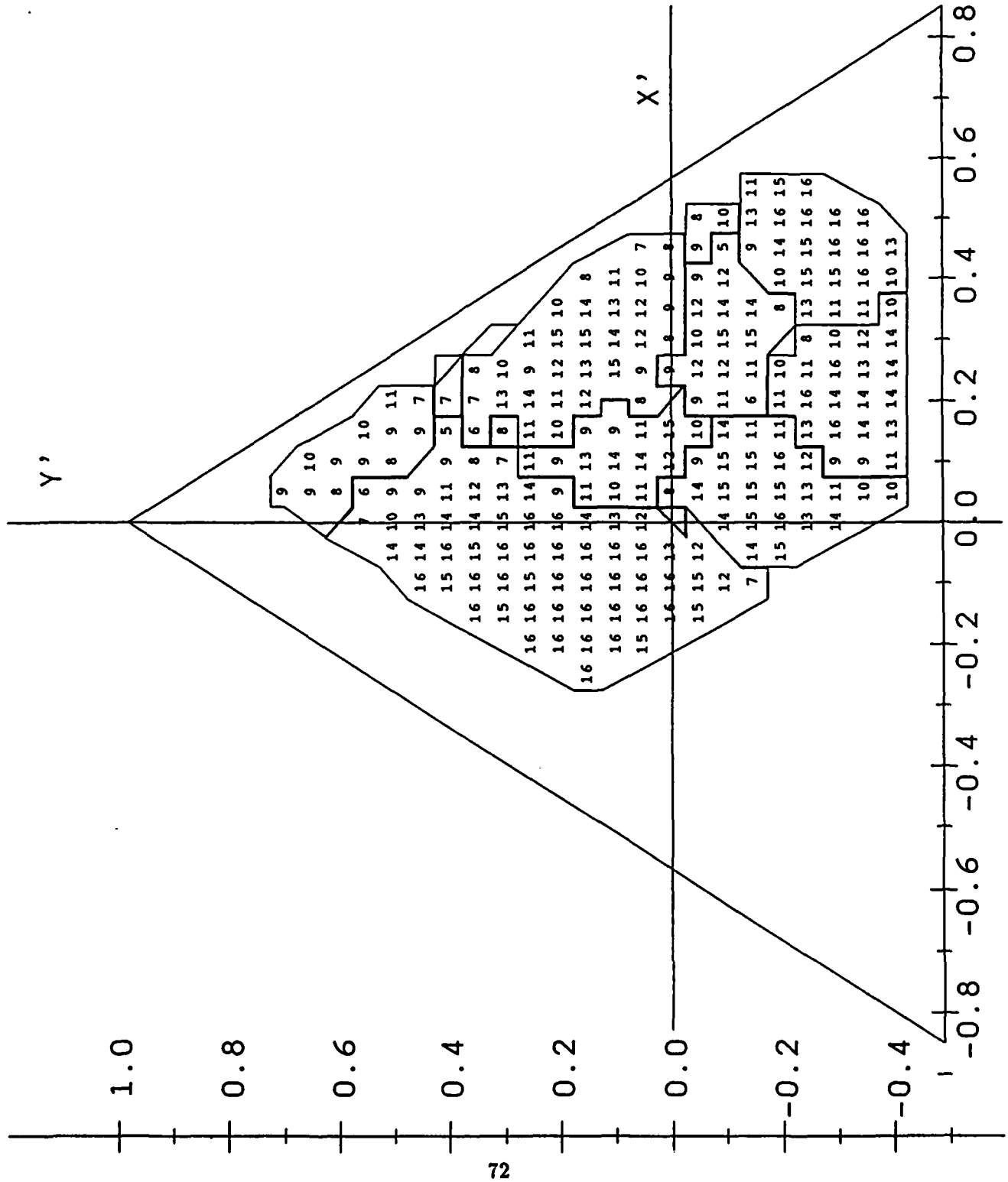


Figure 2-10: (f) Plurality frequencies (See text) for all tokens in the $z' = 0.55$ plane.

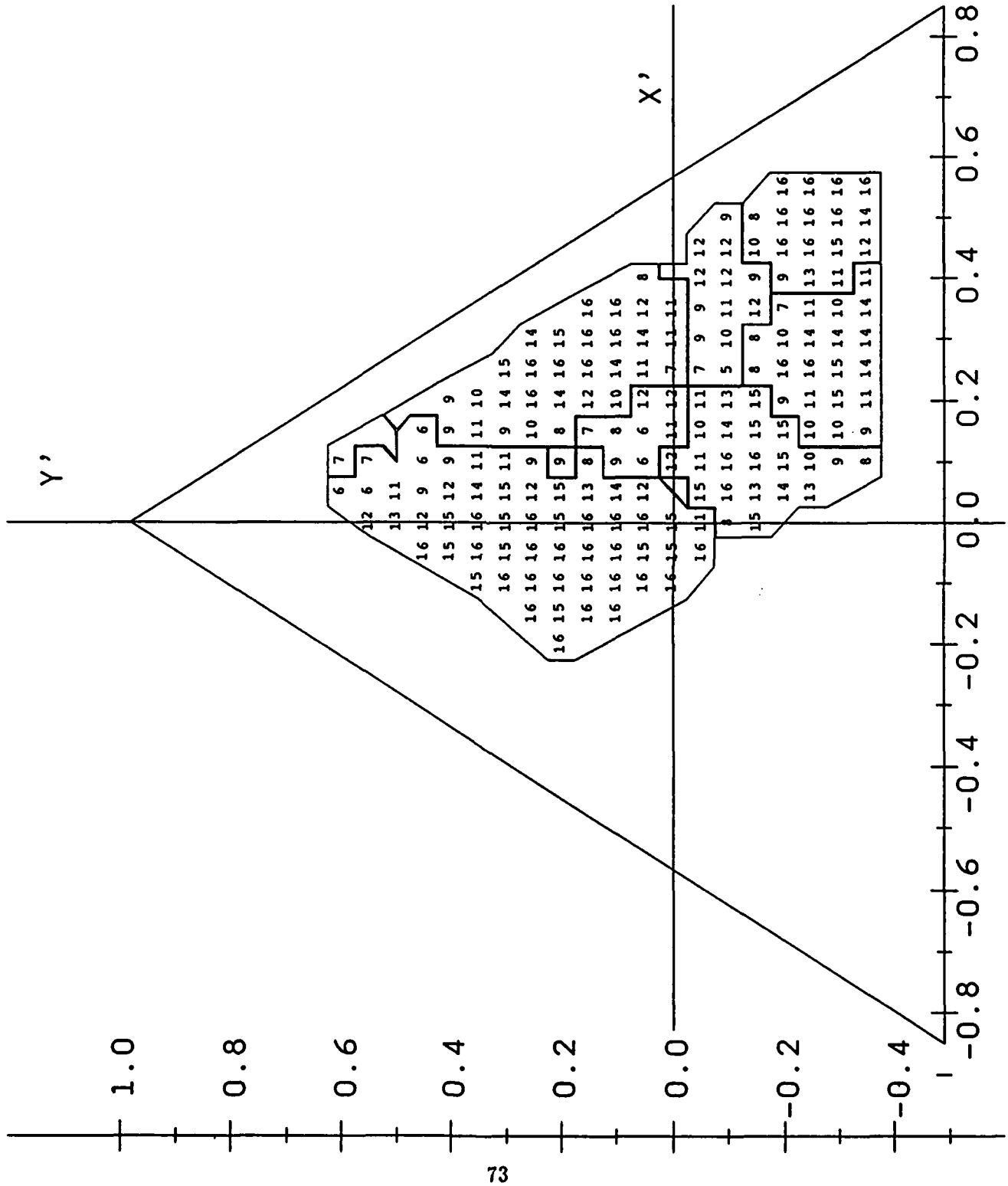
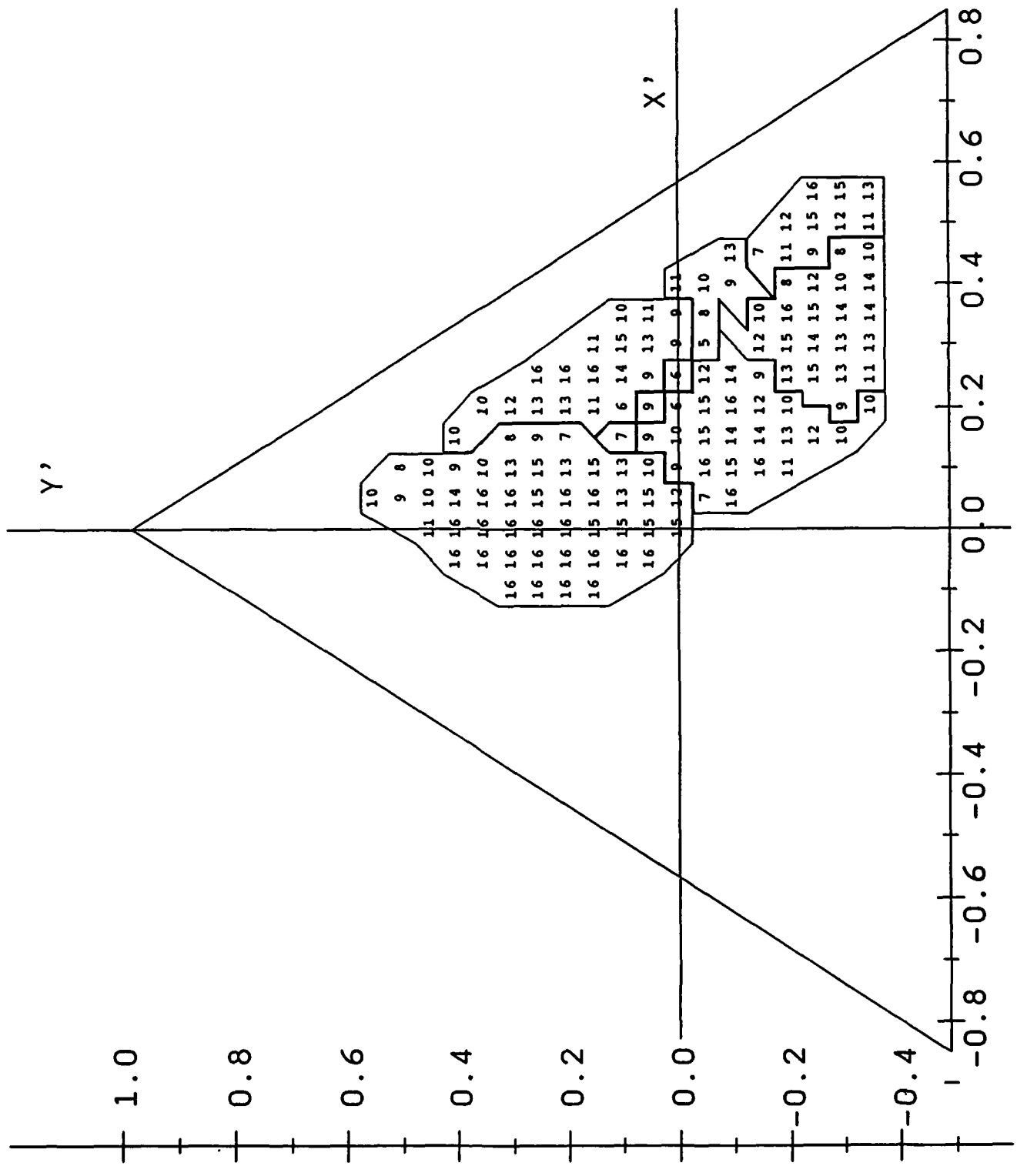


Figure 2-10: (g) Plurality frequencies (See text) for all tokens in the $z' = 0.50$ plane.



are moderately correlated with plurality frequencies, that is, tokens with a greater number of subject responses agreeing on the identification tend to also have higher sums of confidence ratings, the sums of confidence ratings themselves do not appear to substantially add information pertaining to the relative saliency of tokens.

2.3.8 Individual differences in identification responses

As can be noted from Table 2.5, all subject responses agree on only 19.3% of their identifications, or 170 tokens in the primary planes. However, the average within-subject identification agreement, that is, the average of agreements between subjects' first and second response sets, is 75.9% for these planes. This indicates that a subject agrees with himself on the identifications of an average of about 670 tokens. If subject identification consistency is considered an indicator of vowel saliency, then, on the average, 56.6% of the tokens in the primary planes may be considered salient vowels to individual subjects, but yet are classified differently across subjects. In addition, only 304 tokens were shared in common among tokens where within-subjects agreements occurred. Thus, approximately half of the 56.6% represents different tokens to different subjects.

It is acknowledged that some agreements within and between subjects may occur randomly and do not indicate saliency, but rather, add noise to the data. To reduce this noise and pursue this issue further, a third set of identifications for seven of the eight subjects can be considered for one of the primary planes. All subjects (except 1M) classified the tokens in the $z' = 0.70$ plane as part of their initial training. In terms of subject identifications, this may be considered one of the most salient vowel planes in the experiment. Since $F3$ remains constant for any given plane, the response uncertainty for the training set may have been lower than for the actual experiment where tokens from all planes were randomly presented. However, this difference may have been negated by the fact that, at this point in training, the subjects also had less experience with the task.

Percentages of agreement for the identifications from the training response sets and the first and second response sets from the experimental for the $z' = 0.70$ plane were calculated for each of the seven subjects. The average agreement across the seven subjects was 82.7%. The locations of the tokens for which identifications agreed across the three

Figure 2-11: (a) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.80$ plane.

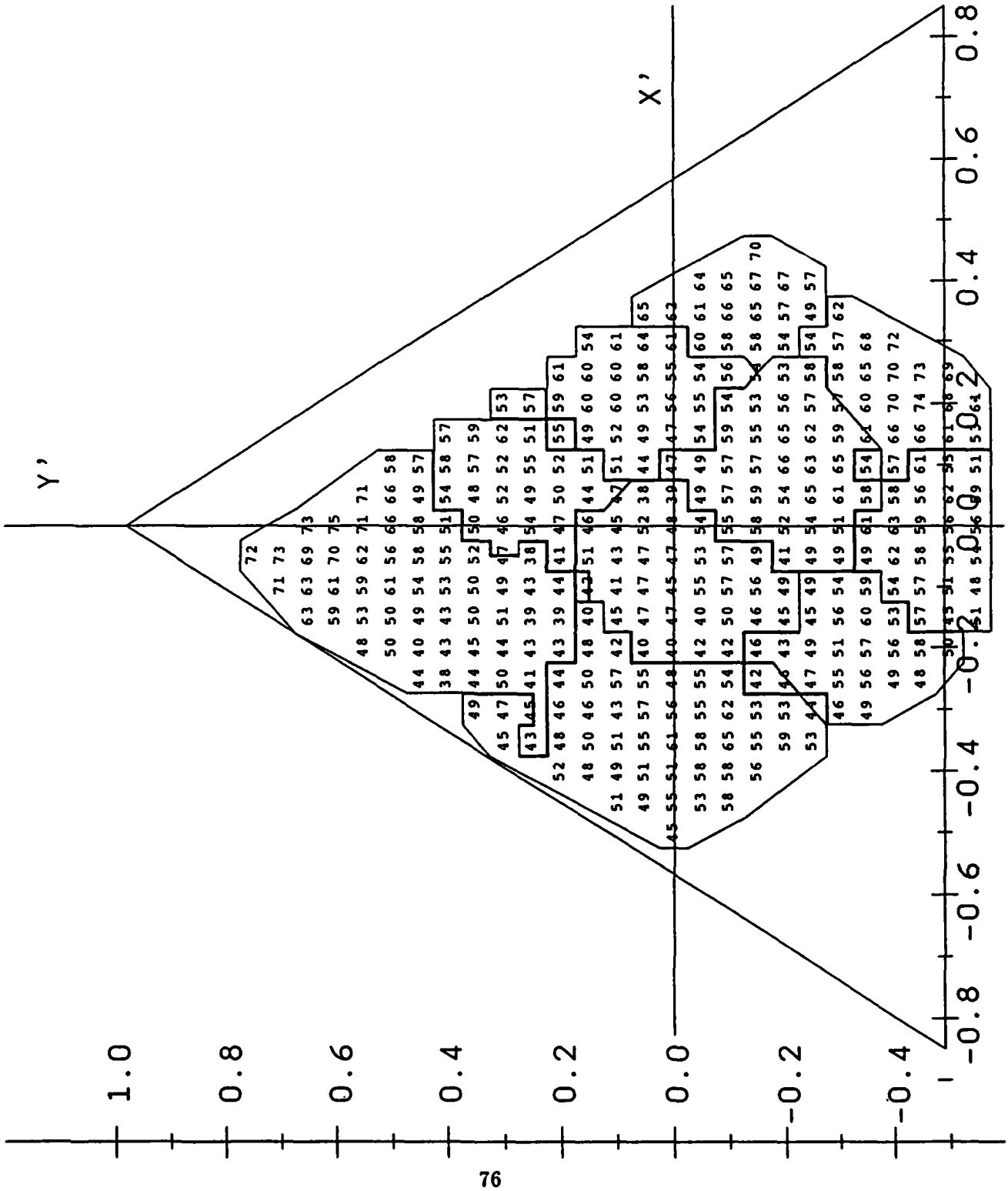


Figure 2-11: (b) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.75$ plane.

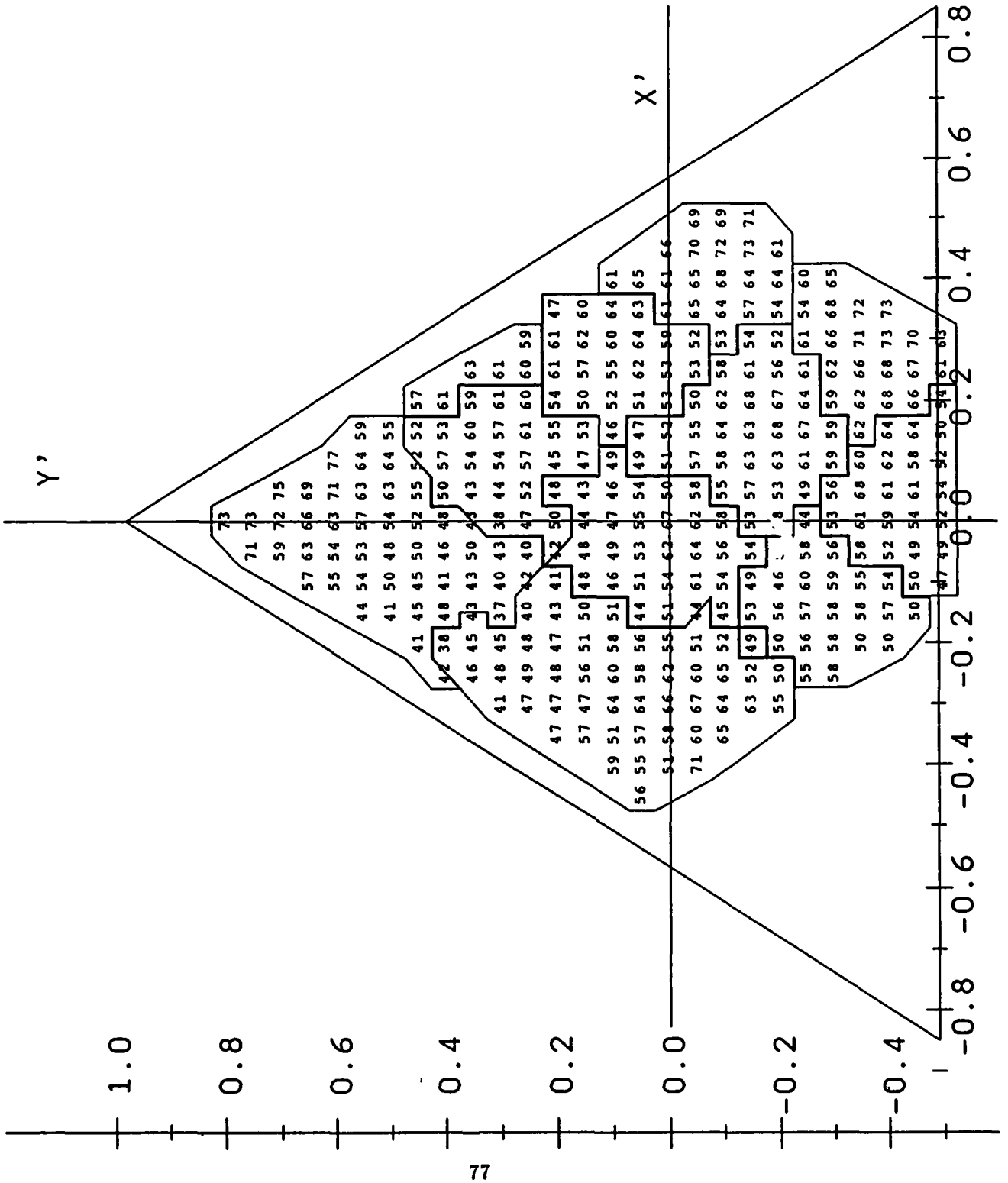


Figure 2-11: (c) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.70$ plane.

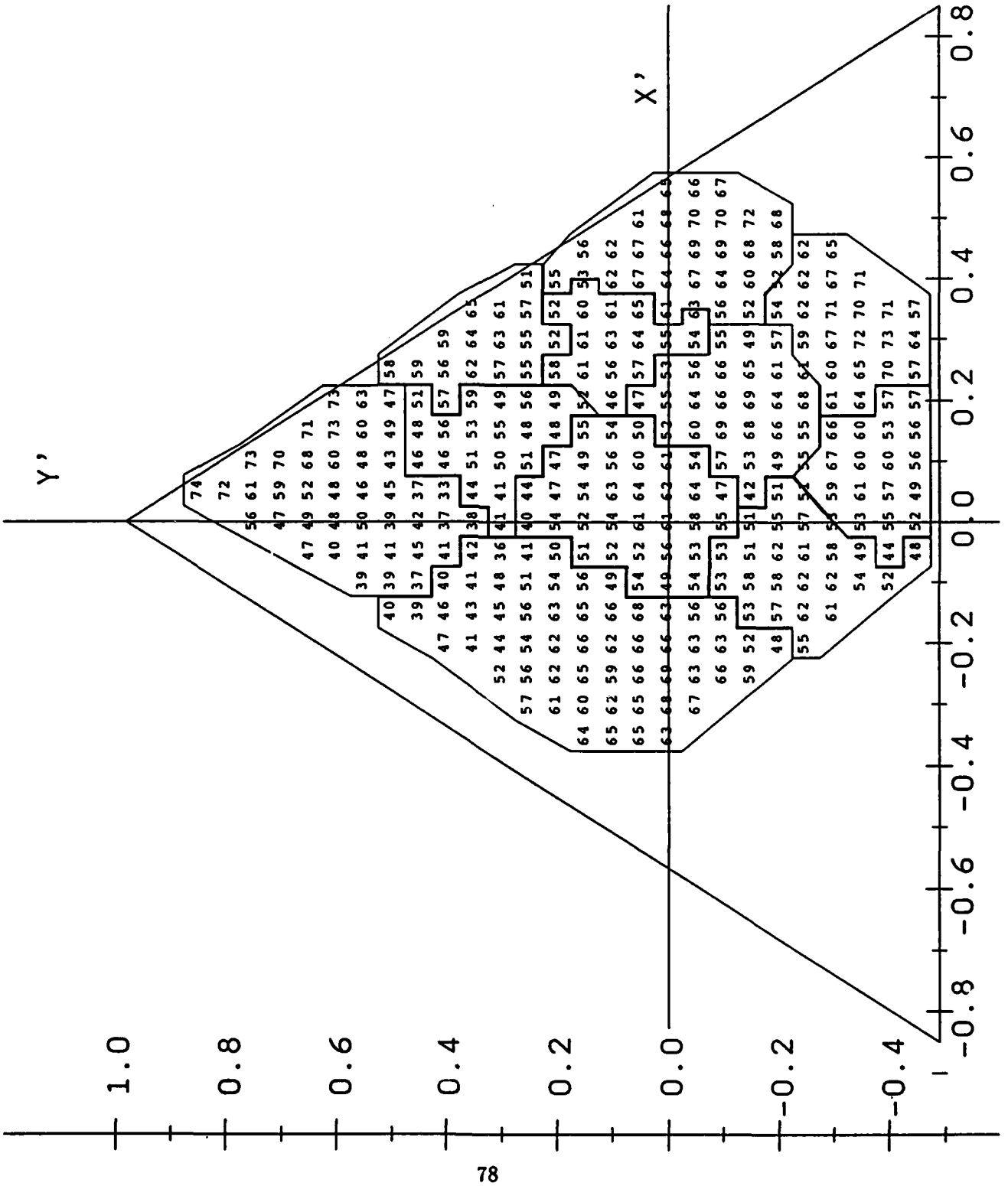


Figure 2-11: (d) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.65$ plane.

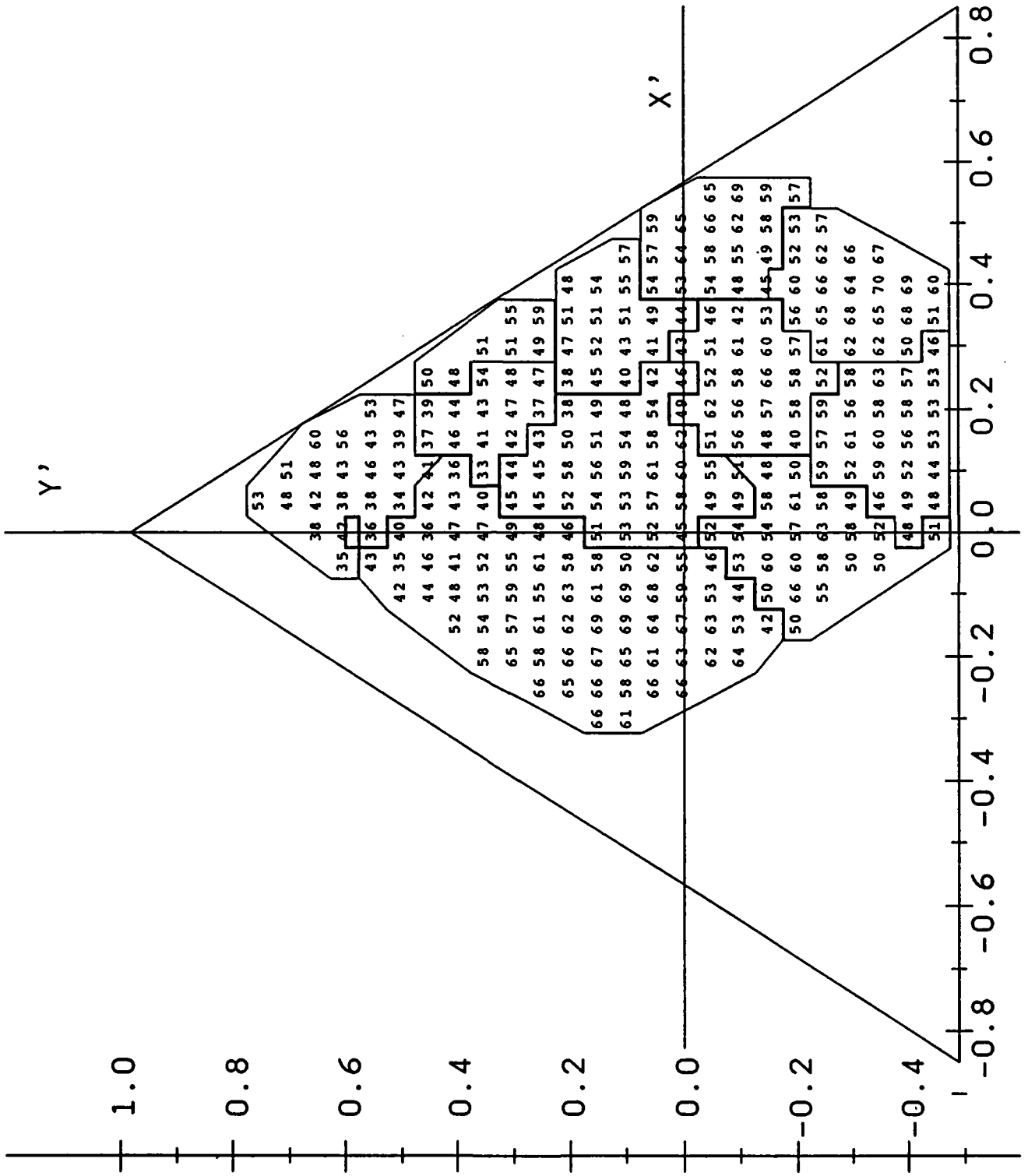


Figure 2-11: (e) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.60$ plane.

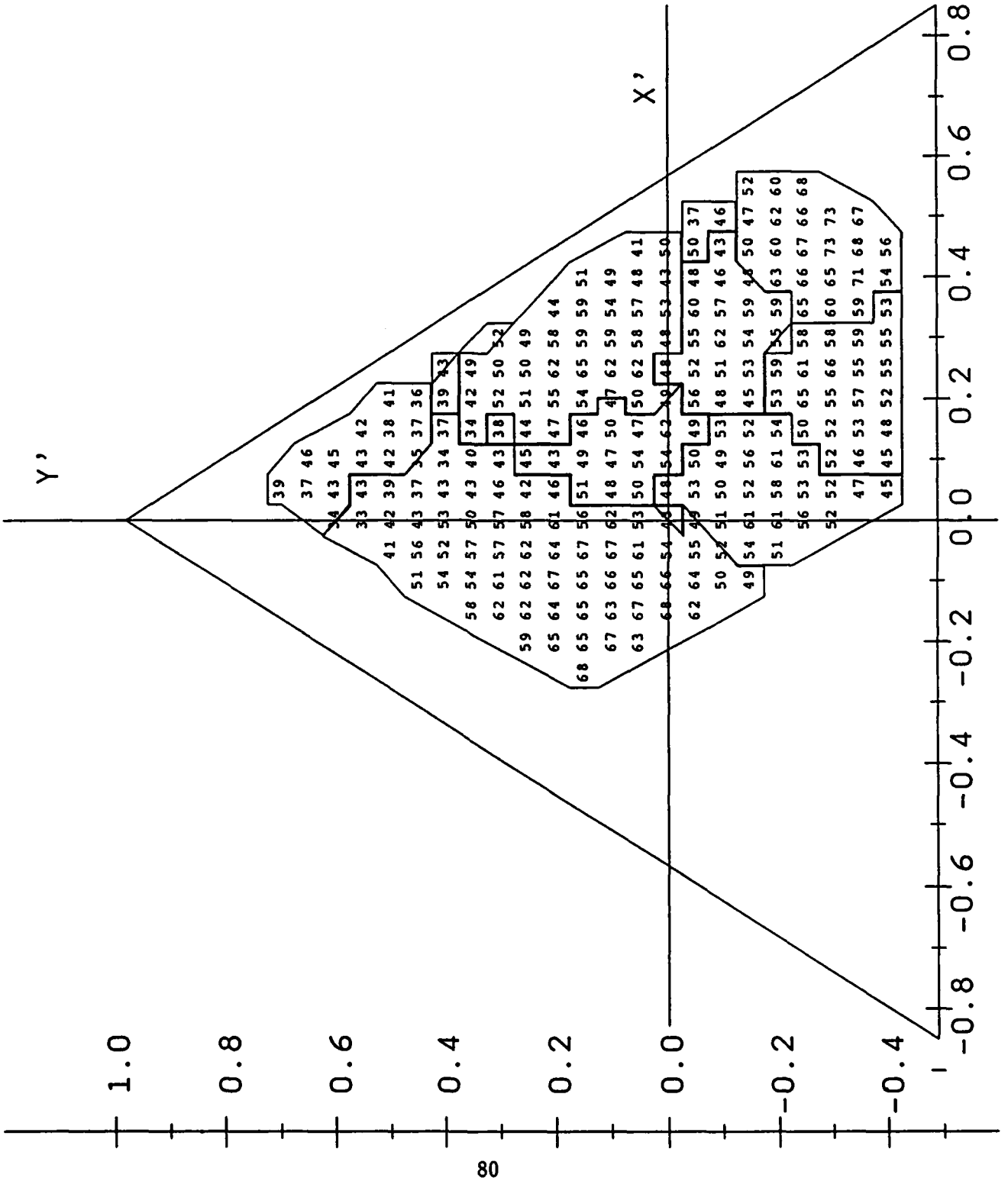


Figure 2-11: (f) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.55$ plane.

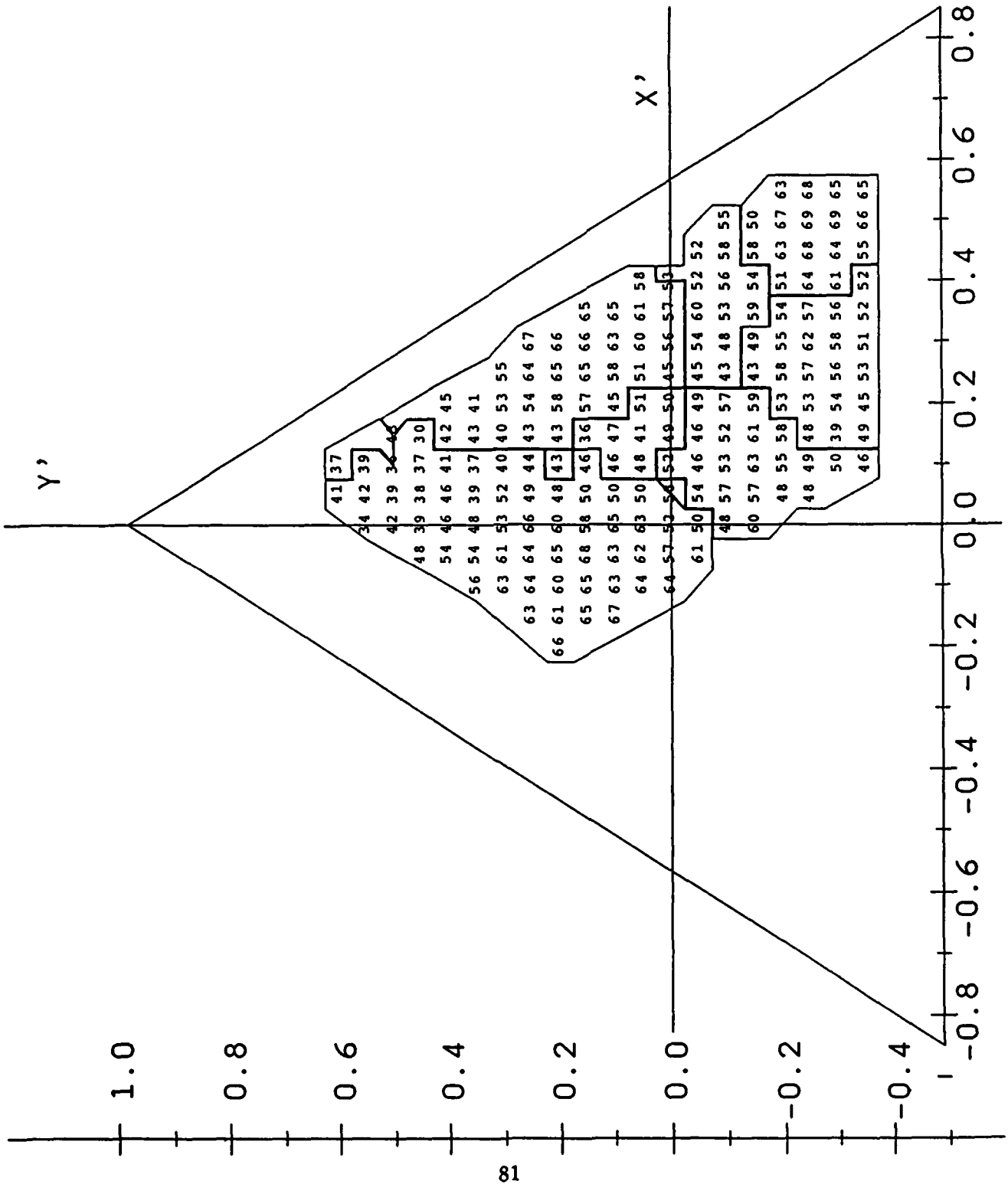


Figure 2-11: (g) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.50$ plane.

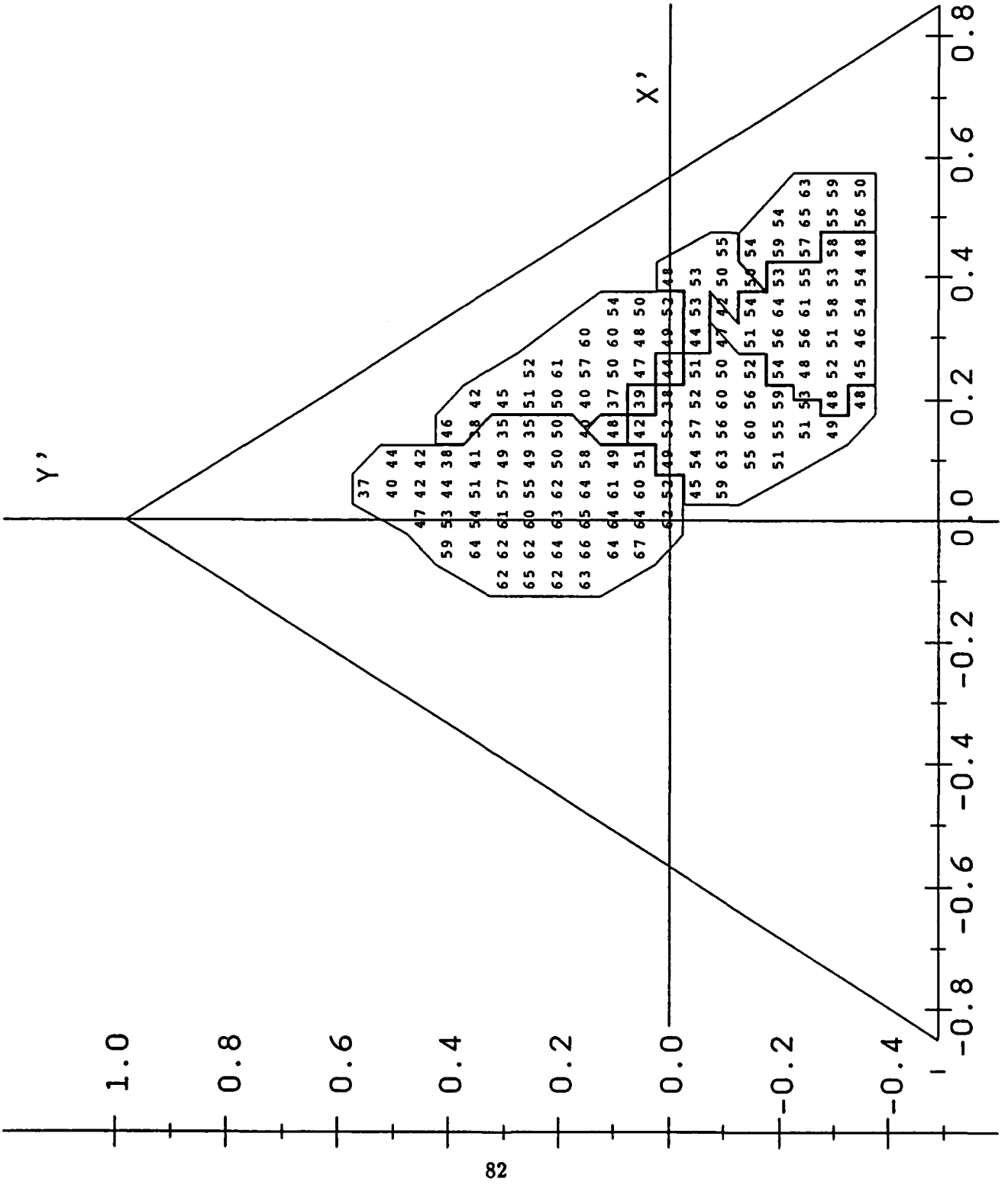
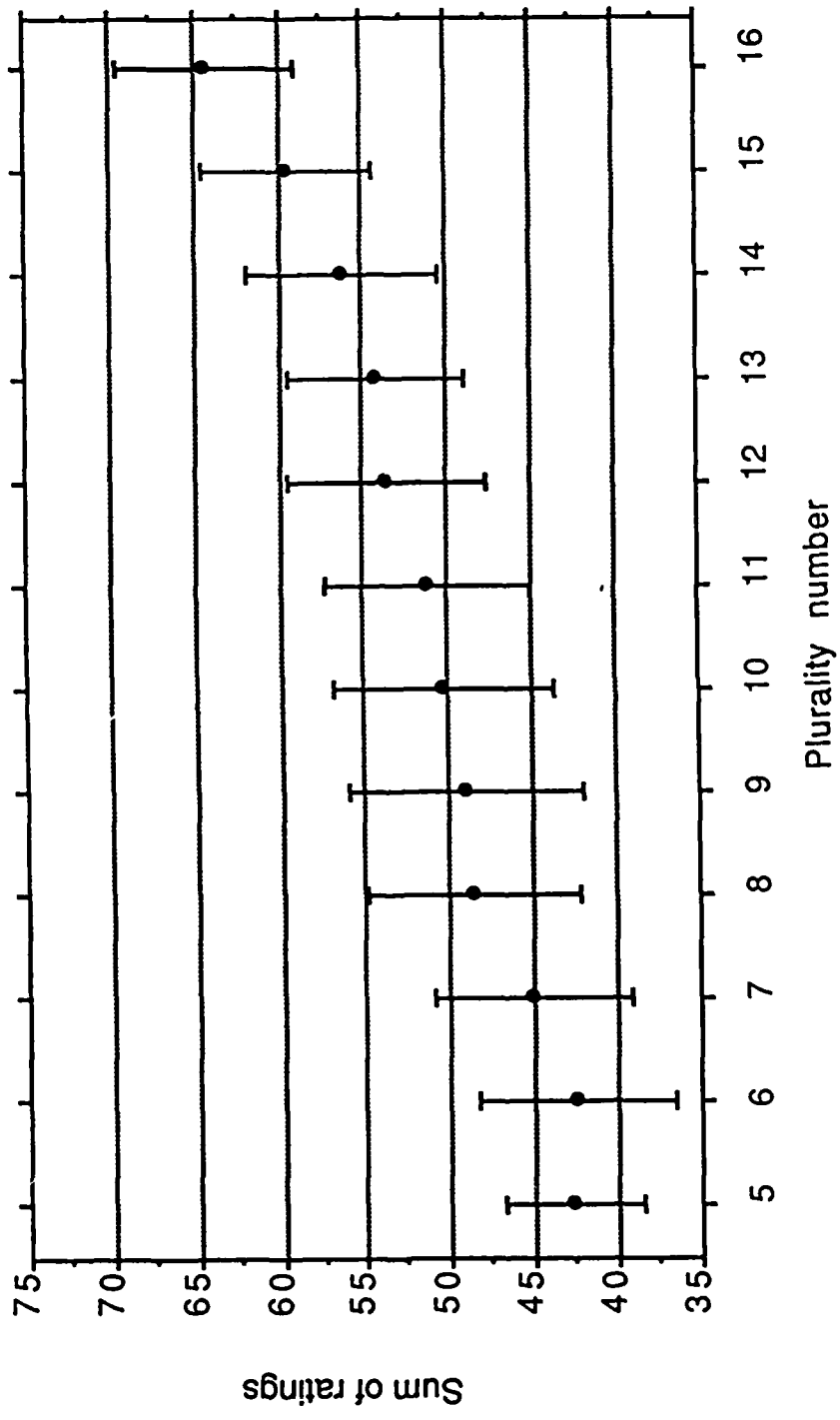


Figure 2-12: Mean sums of confidence ratings for tokens grouped by plurality frequency. Error bars indicate ±1 standard deviation.



response sets are shown in Figures 2-13a-g for each subject along with the *SSB* target zones for the $z' = 0.70$ plane. In these figures the identifications encircled either disagree with the identifications of the zones in which they fall or fall on boundary lines between zones. These encircled identifications should represent tokens of high saliency to individual subjects, although identified differently or ambiguously by the plurality of all subjects. While a number of the encircled points fall in areas noted previously (see Section 2.2.1) as seemingly difficult for subjects to identify (the areas at the /UW/-/IY/ and /IH/-/EY/ borders) and a few appear to be obvious errors (note the /EY/ responses in the /OW/ and /AO/ zones in Figure 2-13g), the majority fall near or on boundary lines. This analysis suggests that certain formant combinations may produce very salient perceptions of a given vowel quality to individuals which are different from the perceptions of a majority of individuals. Such a suggestion supports the notion that vowel perception may be greatly influenced by individual differences when stimuli are at or near a generalized boundary between two vowel categories. This will potentially add difficulty to defining an accurate general predictive model of vowel perception if based solely on the acoustic attributes of vowels.

Figure 2-13: (a) Locations of tokens for which identifications agreed across three response sets for subject 1F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

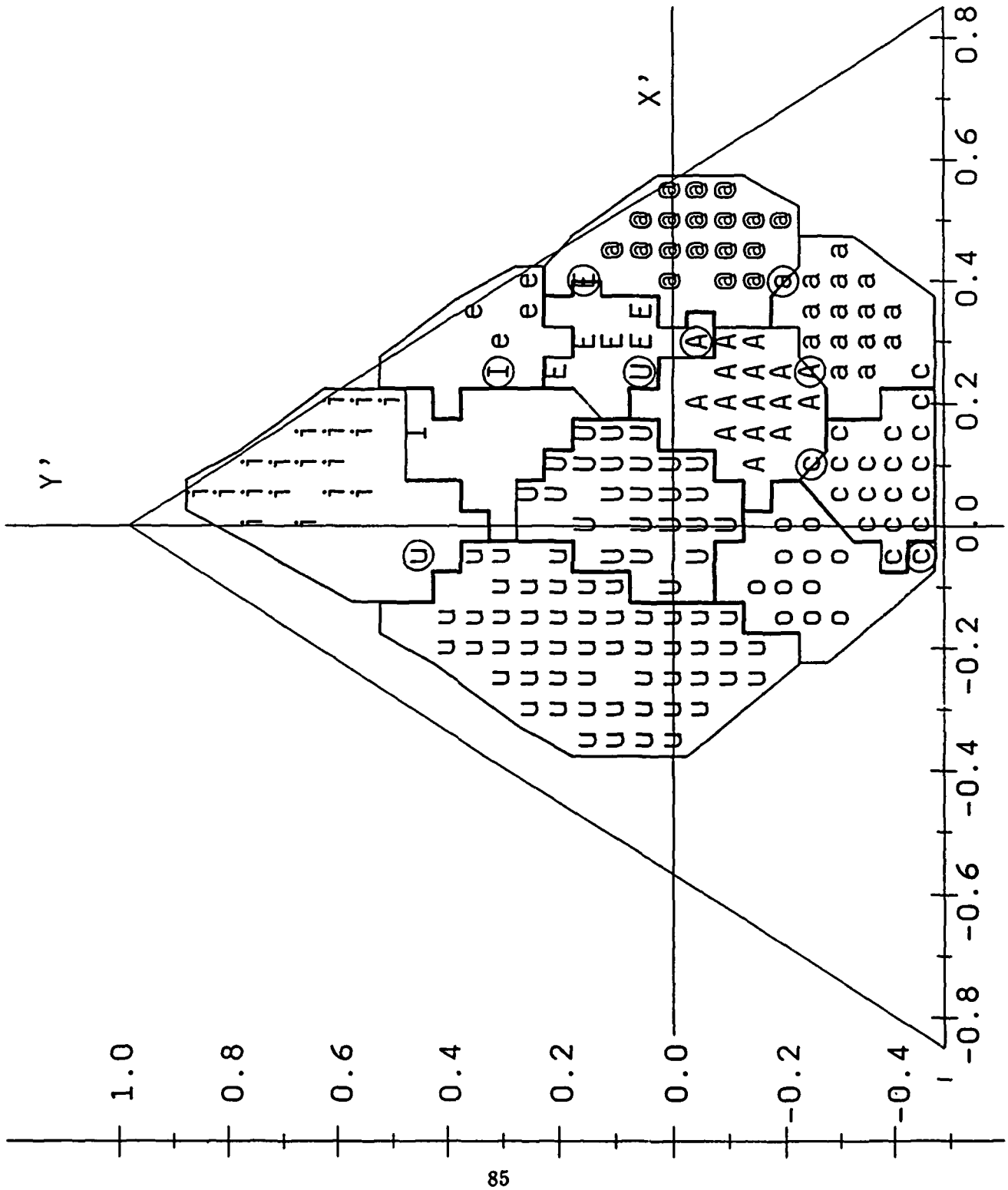


Figure 2-13: (b) Locations of tokens for which identifications agreed across three response sets for subject 2F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

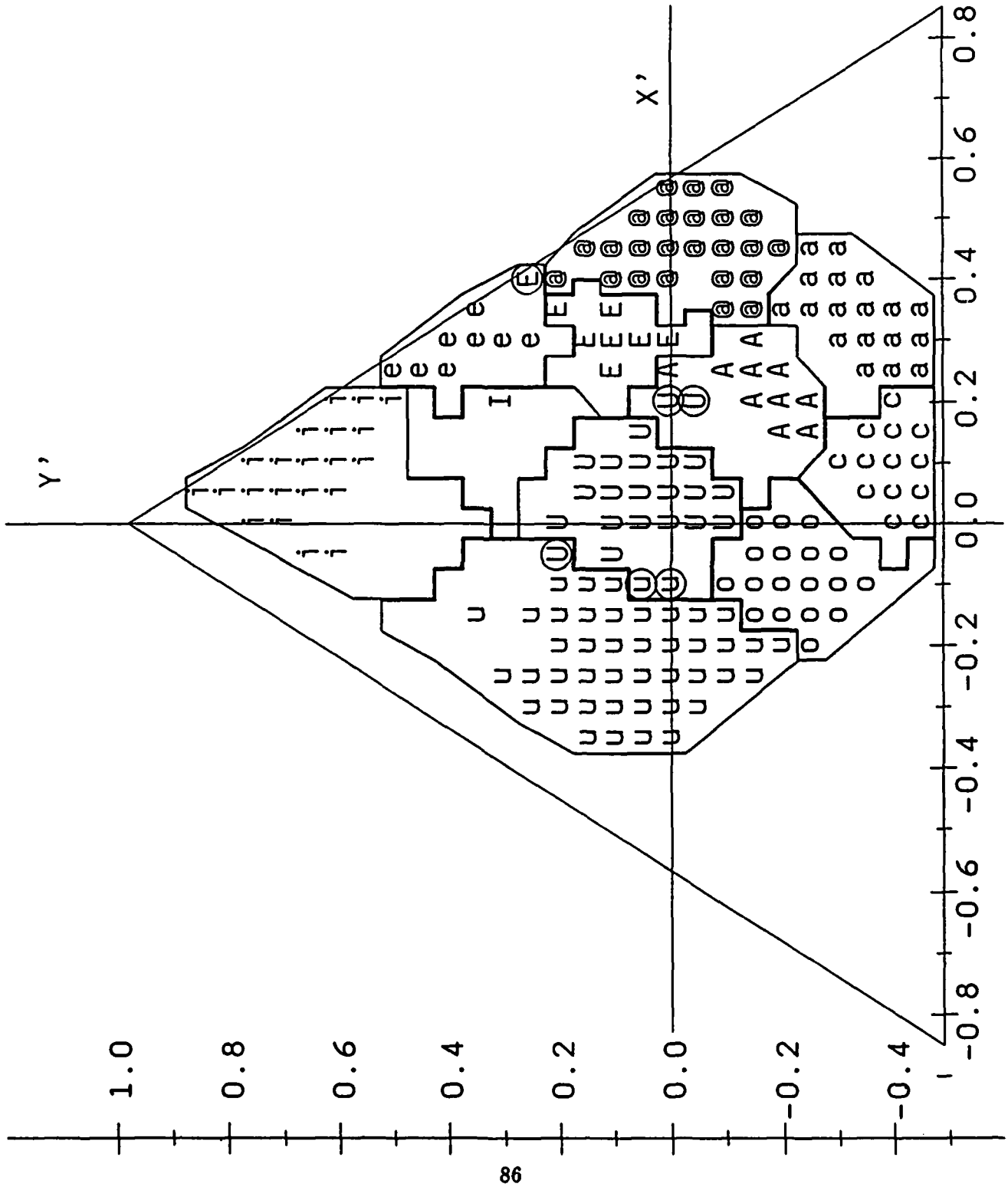


Figure 2-13: (c) Locations of tokens for which identifications agreed across three response sets for subject 3F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

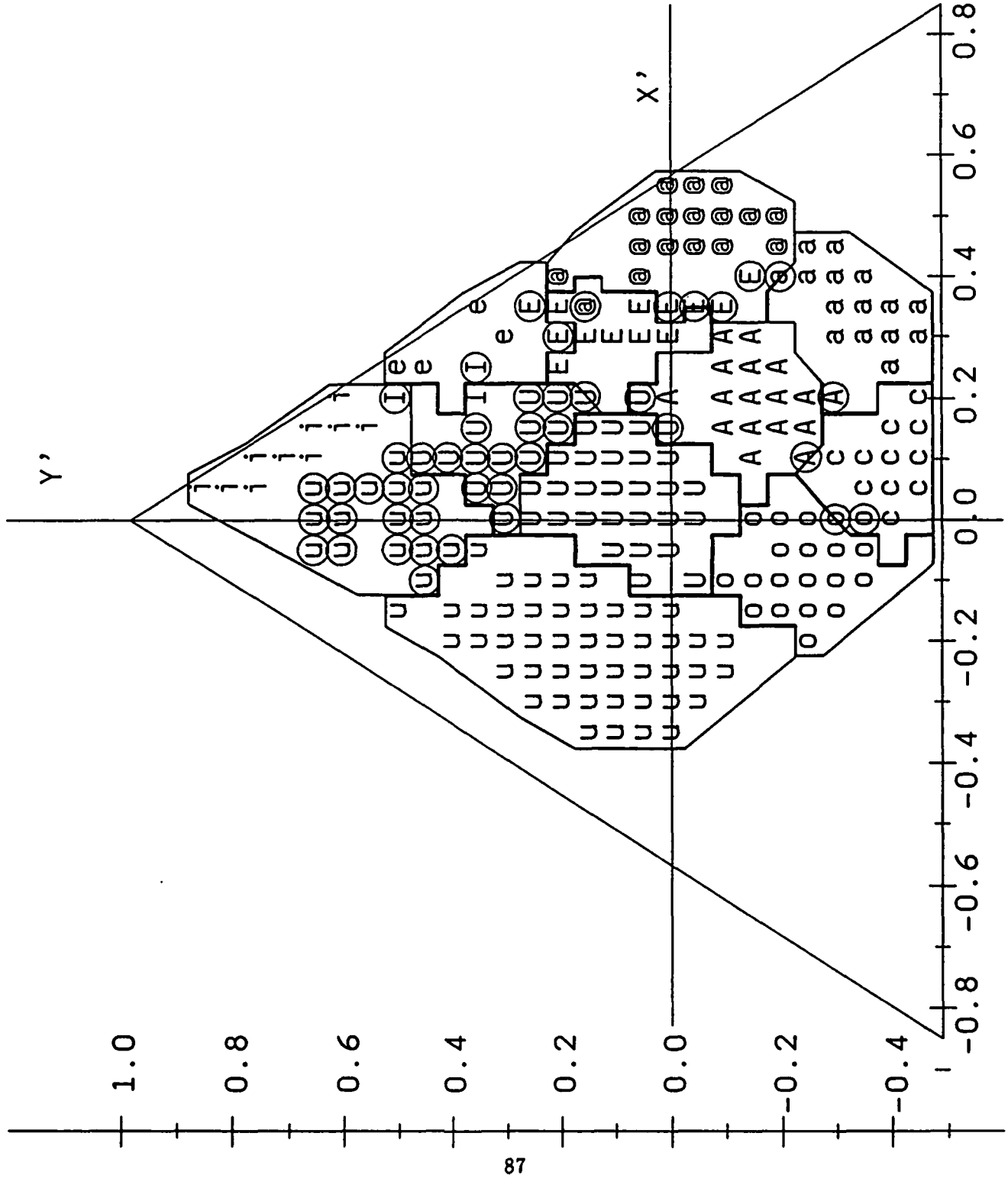


Figure 2-13: (d) Locations of tokens for which identifications agreed across three response sets for subject 5F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

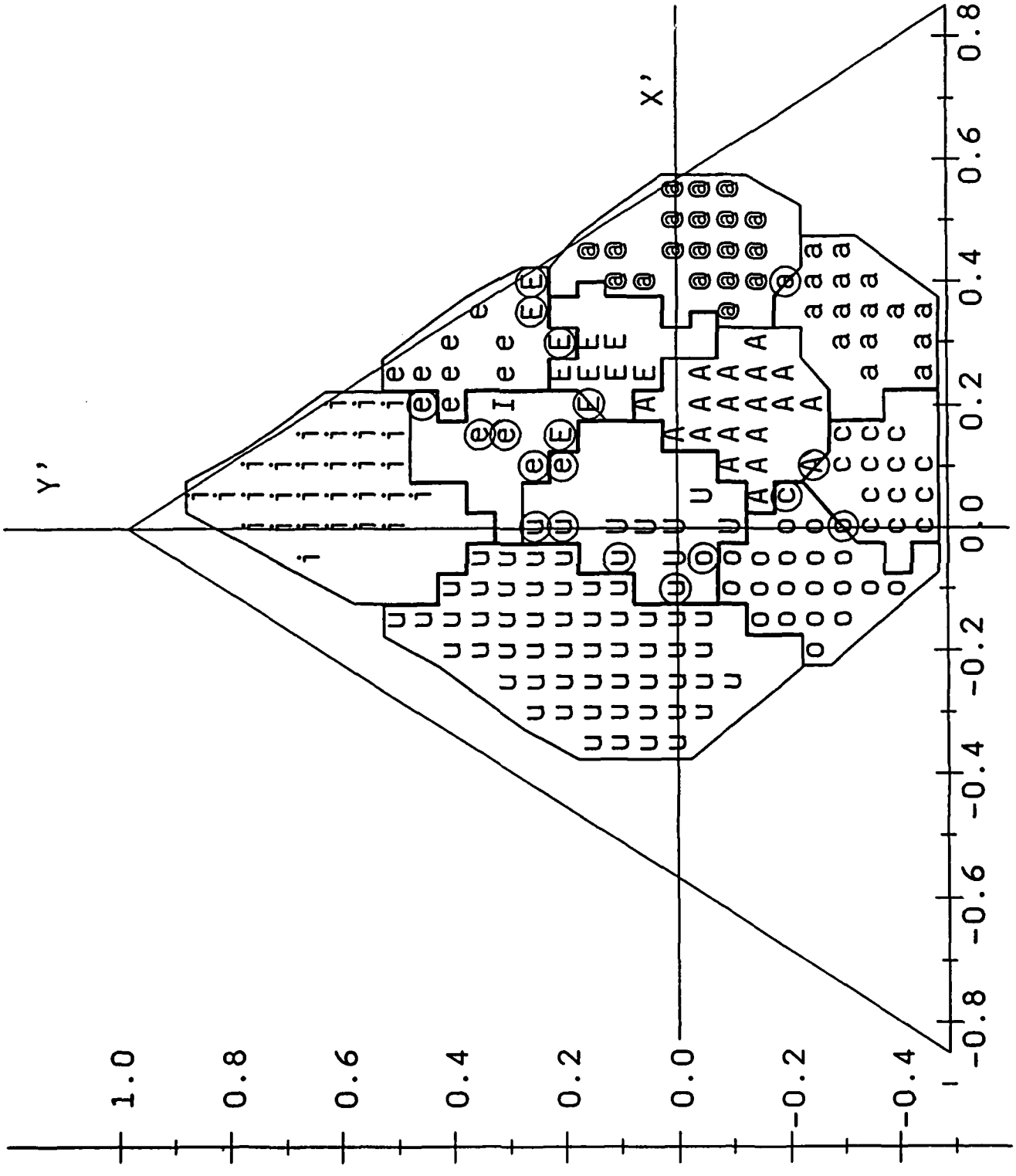


Figure 2-13: (e) Locations of tokens for which identifications agreed across three response sets for subject 3M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

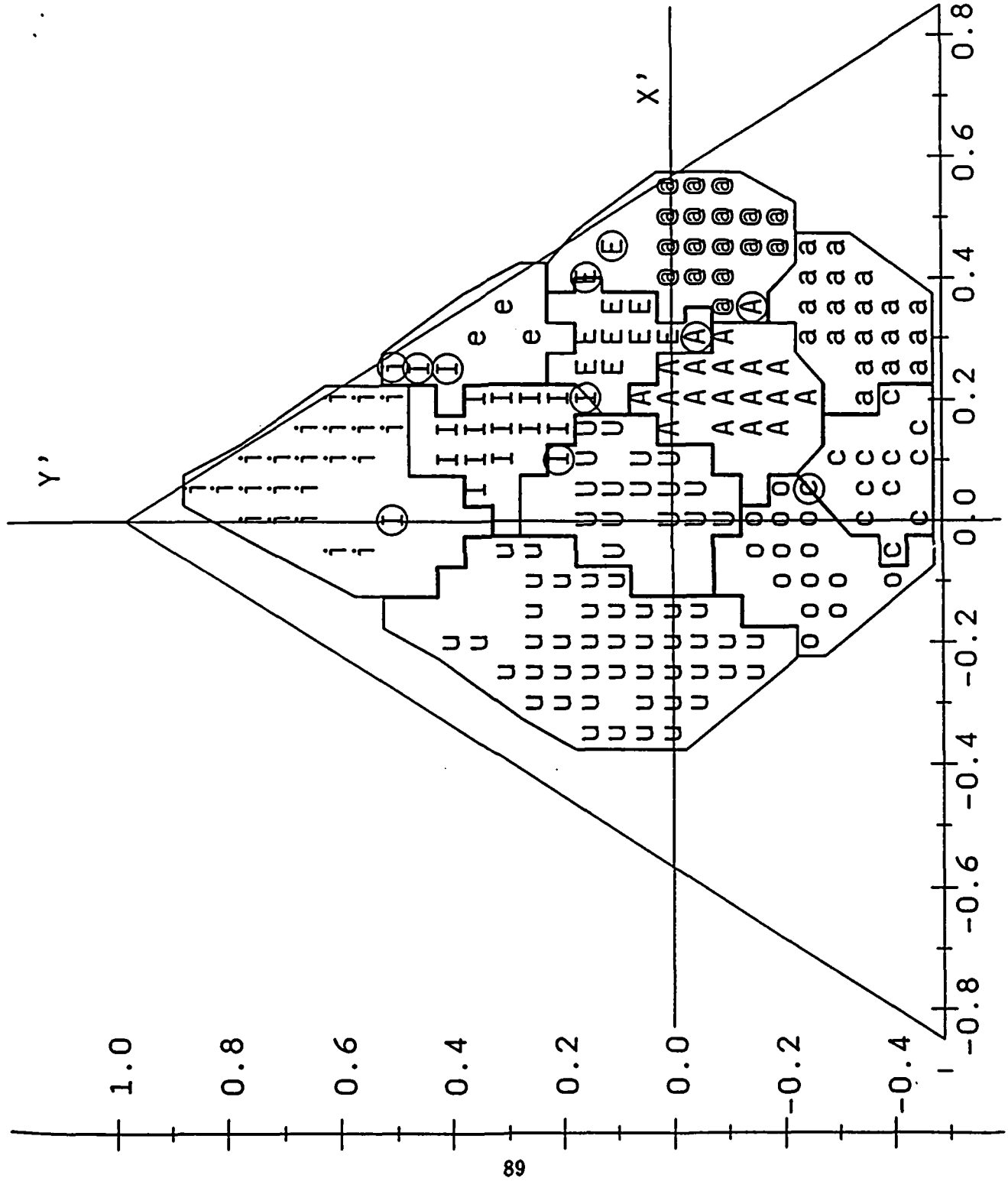


Figure 2-13: (f) Locations of tokens for which identifications agreed across three response sets for subject 4M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

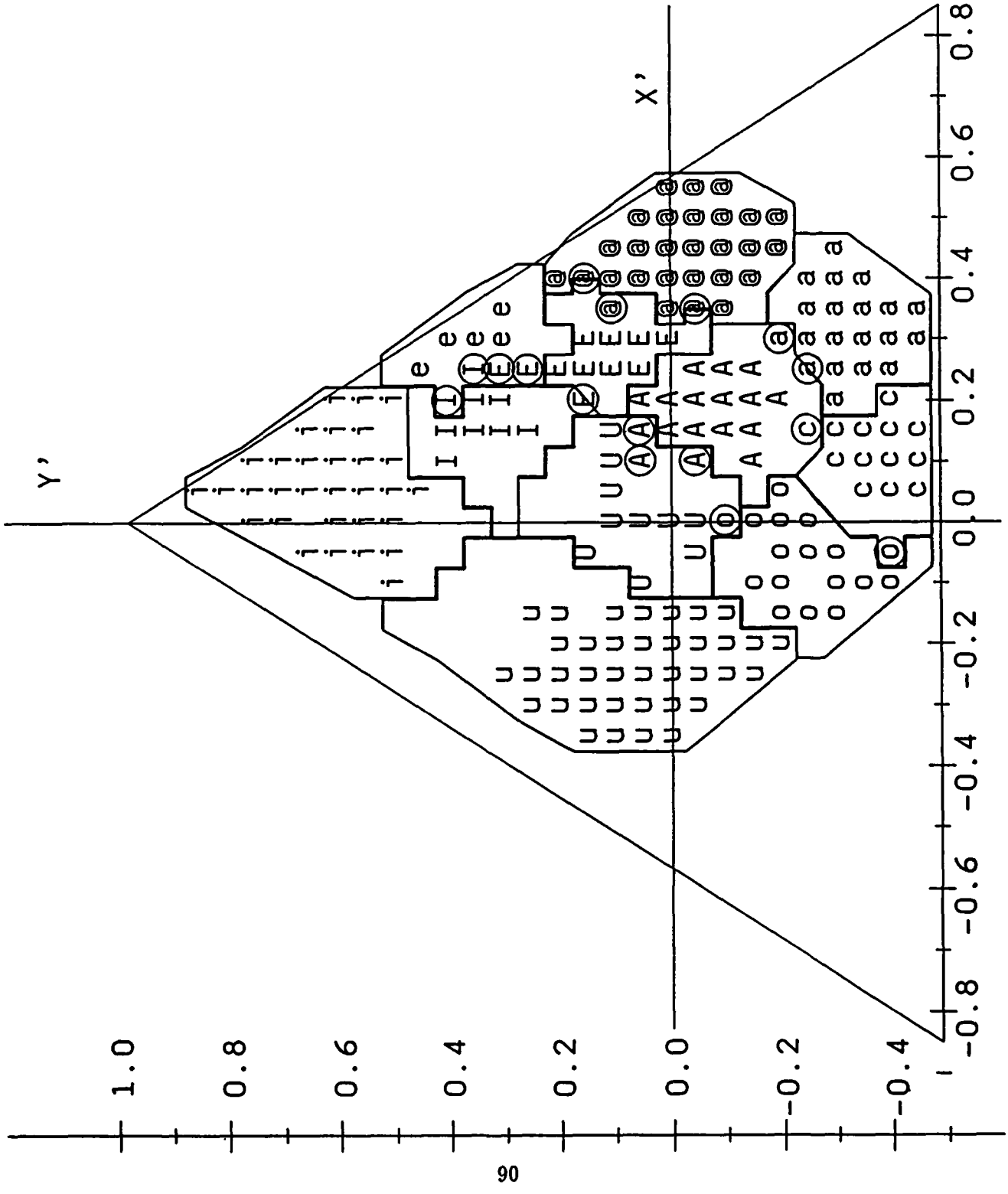
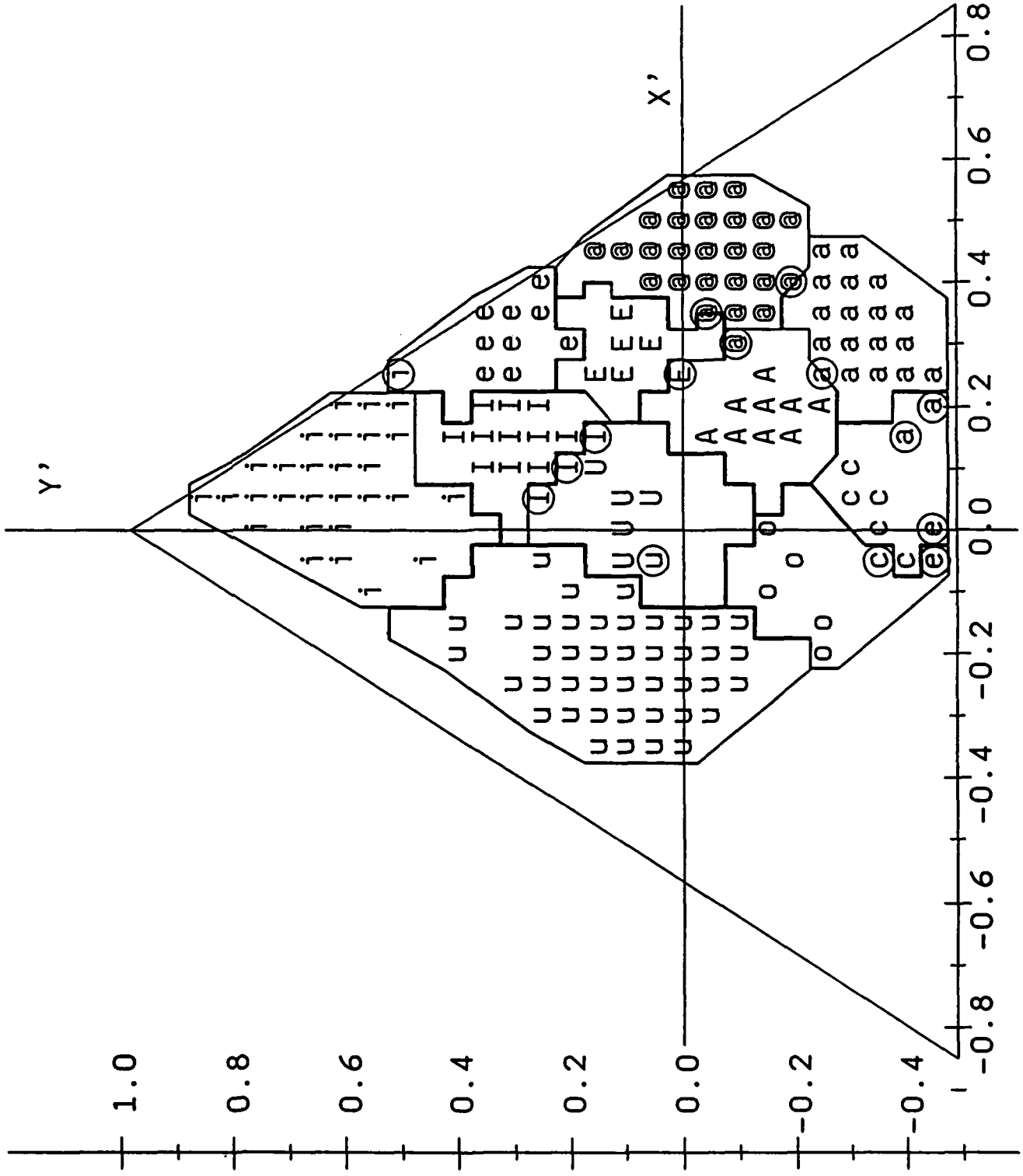


Figure 2-13: (g) Locations of tokens for which identifications agreed across three response sets for subject 5M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.



2.3.9 Linear Discriminant Analyses

To investigate the relation between the plurality identifications of the synthetic vowel tokens and their acoustic variables, a series of linear discriminant function analyses (Tatsuoka, 1970) were performed on the data. This analysis technique has often been utilized in past speech research (for examples, see Neary, 1977; Assmann, Neary, and Hogan, 1982; Syrdal and Gopal, 1986) to provide indices of resolution. These analyses perform a statistical classification function utilizing continuous variables along one or several dimensions, and base the classification on an *a posteriori* probability (*APP*) of group membership assuming multivariate normal distributions of the variables about group means. Results will be presented here using the *R* (resubstitution) method of classification whereby each token is classified according to calculations based on all the data. The percent correct classification calculated by this method provides a relative index of resolution for how well a given data set may be divided into groups, and the average *APP*, an index of the relative strength of group membership, where group means become more widely separated and group data more closely clustered as the average *APP* approaches one. Thus, linear discriminant analysis provides a method for quantitatively describing how a set of variables may classify a given set of data in a linear statistical fashion and will be used here to evaluate how traditional vowel measures (i.e., $F1$, $F2$, and $F3$), as well as *APS* coordinate values, account for the plurality identifications.

Table 2.6 shows the percent correct and average *APP* scores from a series of linear discriminant function analyses on the 1674 plurality identifications (excluding rejected points) with respect to various combinations of $F1$, $F2$, and $F3$; x, y, and z; and x', y', and z'. It is apparent from the table that classification by $F1$ and $F2$ is quite accurate, but that classification accuracy does increase when $F3$ is included. Classification by log ratios of formants appears inferior to simple formants until ratios representing all formant information ($F1$, $F2$, and $F3$) and $F0$ is utilized. At that point, accuracy is approximately equal to simple formant performance. The performances of x, y, z and x', y', z' are virtually identical since both variable sets represent the same four parameters. However, the x', y' combination yields somewhat higher accuracy than the simple log ratios because $F0$, $F1$, $F2$, and $F3$ are all still represented in various proportions by these two coordinates.

Table 2.6: Linear discriminant analyses of plurality identifications.

Variable(s)	% Correct	Average *APP*
$F1, F2$	80.8	0.671
$F1, F2, F3$	85.3	0.700
$z : log(F2/F1)$	37.6	0.275
$y : log(F1/SR)$, z	81.6	0.694
$x : log(F3/F2)$, z	48.6	0.365
x, y, z	85.6	0.726
x', y'	61.5	0.496
x', y', z'	85.7	0.726

If the utilization of $F3$ in vowel perception is predominantly that of a retroflexion detector, then the addition of $F3$ to $F1$ and $F2$ alone in accounting for the data should be minimal when /ER/ identifications are removed from the data set. The results of linear discriminant analyses for such a data set are shown in Table 2.7. Note that not only are all indices higher, but the addition of $F3$ to $F1$ and $F2$ alone increases the percent correctly classified by only 0.4%. This suggests that indeed $F3$ may primarily be used by subjects to mediate the percept of retroflexion and that $F1$ and $F2$ are the primary attributes used to mediate non-retroflex vowels.

Besides providing an index of the relative group membership strength, reflected in the average *APP*, linear discrimination analysis can provide the *a posteriori* probability of group membership for individual tokens. Assmann, Nearey, and Hogan (1982) attempted to predict vowel identification responses by examining the correlations of *APP* scores for individual tokens from linear discriminant analyses with the identification rates of 100 natural isolated vowels and gated portions of the same vowels. They found correlations, using the Spearman rank order correlation statistic, ranging from $R^2 = 0.007$ to 0.490, depending on the analysis variables and data set manipulations under test. They cited four reasons for considering this degree of correspondence to be noteworthy: 1) the *APP* scores were based on a small vowel sample compared to the listeners' presumably larger experience

Table 2.7: Linear discriminant analyses of plurality identifications (no /ER/).

Variable(s)	% Correct	Average *APP*
$F1, F2$	87.0	0.729
$F1, F2, F3$	87.4	0.734
x', y'	61.6	0.494
x', y', z'	87.1	0.755

base; 2) the acoustic variables used in the analyses may not have been those used by the perceptual system; 3) errors in measurement could not be ruled out; and, 4) context effects may have been present which can influence vowel identification.

If a sufficiently high correlation were found between *APP* scores for individual tokens and identification plurality rates for the current data, the *APP* scores could serve as a powerful predictive tool for establishing a saliency gradient, as was discussed in sections 2.3.6 and 2.3.7, for the vowel zones in *APS*. However, a correlation higher than those found by Assmann and colleagues would be required to serve any useful purpose.

The *APP* score for the classification of each token (except rejected points) was first obtained from the linear discriminant analysis yielding the highest index values (x', y', z'). A Spearman rank order correlation was then computed between the plurality frequencies expressed as a proportion of the total possible number of responses (16) and the corresponding *APP* scores of all tokens. A moderately low correlation (R^2= 0.381) was found after correction for ties. Although this low correlation suggests that the *APP* scores are not adequate for use in a predictive model for identification rates, it does agree well with the correlation (R^2= 0.373) found by Assmann et al. (1982) for identification rates of isolated vowels presented in a mixed speaker condition and *APP* scores generated using natural log-transformed values of $F1$, $F2$, and $F3$.

In summary, linear discriminant analysis is a statistical classification tool capable of providing relative indices of resolution and group membership strength based on *a posteriori* knowledge of the distributions of the variables under test. Analyses of this type on the current data suggest that a relatively high level of correct classification for vowel identifications

can be achieved utilizing $F1$ and $F2$ as variables and that the addition of $F3$ as a variable provides only minimal improvement except in the classification of the retroflex /ER/. Correct classification utilizing log ratios of formants represented by x, y, and z, or transforms of these variables represented by x', y', and z' is approximately equivalent to classification utilizing the first three formants. While the *a posteriori* probabilities for individual tokens have been suggested as possible predictors of identification rates, correlation between these probabilities and pluralities for the present data are low, suggesting that such a strategy is unsuitable as a predictive model.

2.3.10 Agreement by z' plane

As has been mentioned previously, movement along z' in the *APS* primarily reflects a change in $F3$ when *SR* is fixed, such that $F3$ increases as z' increases. Although the data presented thus far suggests that $F3$ is not influential in the perception of non-retrofex vowels, it is of interest to determine whether or not $F3$ plays at least some role beyond the retroflex/non-retroflex distinction in determining the perceptual salience of vowels.

To investigate this issue, the minimum, maximum, and average z' coordinate values were found for all vowel measurements in each of ten vowel categories from the CID natural speech database and from Peterson and Barney (1952). These values, along with the z' range expressed in log units, are shown in Table 2.8 for all vowel categories and the averages of these values across all non-retroflex vowel categories. The data from Table 2.8 indicates that for natural, non-retroflex vowels the highest minimum and maximum values of z' are found for the high front vowel /IY/ and gradually decrease as the tongue position moves down, back, and up again, with the lowest z' values found for the high back vowel /UW/. The average z' values however do not reflect the gradual slope of the minimum and maximum values, staying relatively constant around $z' = 0.70$ except for the extreme high vowels. The retroflex vowel /ER/ has the lowest minimum, maximum, and average z' values of all the vowel categories considered. The range of z' values appears to be somewhat smaller for front vowels and larger for back vowels, with an average z' range of .141 log units.

The average percentages of agreement by z' plane for all vowel categories were calculated from pairwise comparisons of all subject response sets and are shown in Table 2.9. The value

of $F3$ corresponding to each z' plane is shown in parentheses. To normalize across the

Table 2.8: Averaged values of minimum, maximum, and average z' for CID Natural Speech Database and results from Peterson and Barney (1952).

Vowel Category	z' min	z' max	$\overline{x}$ z'	z' range
IY	.679	.796	.729	.117
IH	.657	.790	.708	.133
EH	.655	.767	.705	.112
AE	.645	.784	.700	.139
AA	.622	.777	.703	.155
AH	.633	.774	.699	.141
AO	.620	.783	.699	.163
UH	.622	.773	.684	.151
UW	.612	.768	.671	.156
ER	.531	.727	.596	.196
$\overline{x}$ (No ER)	.647	.764	.700	.141

differences in response frequency of the vowel categories, the agreement percentages for individual vowel categories were calculated as the averaged sum of the agreements between each subject response set pair for that category each divided by the total number of different tokens identified as that category across the response set pair. If the assumption is made that synthetic tokens which reflect formant patterns found in natural speech are more perceptually salient than those tokens which do not, and if subjects' agreement on the identification of a token in some way reflects the salience of the token's vowel quality, we might anticipate higher percentages of agreement for tokens that follow the z' patterns found in natural speech. For the present data, we might expect to find higher percentages of agreement for each vowel category between the minimum and maximum z' values for natural vowels in that category and the highest percentage of agreement for tokens located nearest the z' average. Note that at the most general level, the total average percentages of agreement in Table 2.9 (bottom row) do not differ substantially across z' planes, suggesting that subjects agree with one another about 64% of the time no matter what the value of $F3$. However, comparison of Tables 2.8 and 2.9 at the category level indicates that synthetic

Table 2.9: Average percentages of pair-wise agreements for all subject response sets by z' range.

Vowel Category	z' plane (F3 in Hz) 0.50 (1137)	0.55 (1390)	0.60 (1697)	0.65 (2071)	0.70 (2528)	0.75 (3086)	0.80 (3767)
IY	0.3	6.4	21.1	41.9	52.1	54.3	49.6
IH	0.0	0.5	5.0	12.9	22.7	26.5	18.4
EH	0.0	0.0	8.4	29.1	36.4	39.7	34.1
EY	0.3	0.6	2.8	22.1	38.0	13.2	7.9
AE	0.4	0.5	15.3	59.3	64.8	62.1	51.6
AA	45.8	66.7	62.9	60.6	63.4	65.7	60.3
AH	24.1	29.7	36.7	44.3	49.2	47.5	41.0
AO	47.0	49.2	49.2	49.7	49.7	50.6	49.4
OW	52.5	54.4	57.2	53.0	56.0	57.0	47.6
UH	10.6	18.0	26.8	38.5	36.3	31.3	24.1
UW	72.2	70.7	69.4	67.8	64.2	57.0	47.0
ER	40.6	47.7	35.5	3.7	0.0	0.0	0.0
$\overline{x}$ (all categories)	29.4	31.3	32.5	40.2	48.4	45.9	39.2
$\overline{x}$ (all tokens)	65.5	67.1	63.4	64.0	66.5	64.6	57.7

tokens falling within the approximate $F3$ range found for natural tokens are generally agreed upon to a higher extent than are tokens outside this range. This suggests that tokens having $F3$ values approximating those of natural vowels are more salient than those tokens that do not. While this appears to be true generally, in that the highest total average percentage of agreement for all categories occurs for the $z' = 0.70$ plane, some differences are present at the individual category level which are of interest.

At the individual vowel category level, the data in Table 2.9 appear to be in good agreement with the minimum, maximum, and average z' values found in Table 2.8 for the vowels /IY/, /AE/, /AH/, and /UH/. For the front vowels /IH/ and /EH/, the highest agreement percentages occur in the $z' = 0.75$ plane, higher than the z' average of the natural data for these vowels, and the ranges of better agreement along z' also appear to be shifted up. The mid and low back vowels /AA/ and /AO/ do not appear to reflect the natural data in that the agreements are relatively uniform across all z' planes and no distinct plane of greatest agreement is apparent. Although we have no natural data for the mid vowels /EY/ and /OW/, the mid front vowel /EY/ has its maximum agreement percentage at $z' = 0.70$, but good agreement is over only a very small z' range ($z' = 0.65$ to 0.70), while the mid back vowel /OW/ has more uniform agreements similar to /AA/ and /AO/. The high back vowel /UW/ exhibits a different agreement pattern across z' planes from all other vowels. While relatively good agreement for /UW/ is found for all z' planes, the highest agreement is at $z' = 0.50$, well below the z' average for this vowel category. Agreements for /UW/ then become progressively poorer with increasing values of z'. Agreements for the retroflex vowel /ER/ appear to generally agree with the natural data in range, although the z' plane of maximum agreement is slightly lower than the z' average for the natural data.

If $F3$ plays no role in determining the salience of a vowel beyond the retroflex/non-retroflex distinction, the good agreement between the natural data and the agreement percentages for the front vowel categories may still simply be due to the fact that the values of $F3$ for the planes where $z' = 0.65$ or greater are sufficiently high to accomodate the relatively high $F2$ values required for perception of these vowels. However, if this is the case, we might anticipate finding that all other vowel categories equally salient across z' ranges, since the values of $F2$ associated with their perception should be less than 1137 Hz,

the lowest value of $F3$ utilized in the experiment. While this could be true for the /AA/, /AO/, and /OW/ categories, it does not appear to hold for the other mid and back vowels. Thus, while the identification of a vowel token may be predominantly determined by the values of $F1$ and $F2$, $F3$ may play some role in the perceptual saliency of the token, if saliency and subject agreement are related.

In summary, there is general agreement between the ranges of z' for natural vowels and the ranges of z' for higher average identification agreements of synthetic vowels, although specific differences exist between. This suggests that changes in $F3$ have an effect on the saliency of tokens, even though the identifications for non-retroflex vowels can be well accounted for by $F1$ and $F2$ alone.

2.4 Comparisons of vowel classification schemes

A number of theories of vowel recognition based on acoustic attributes of the speech signal have been developed in the past in an attempt to normalize across inter-and intra-speaker differences and, in so doing, resolve the overlap often found between vowels of differing quality when they are grouped along various acoustic dimensions. These theories have sometimes been categorized into two groups determined by the type of information postulated as utilized by the listener in vowel perception (Ainsworth, 1975; Neary, 1989). Theories requiring *extrinsic* specification assume that information from a number of vowels of a single talker are utilized to establish a reference for perception. Theories of this kind would seem difficult to utilize for classification of some synthetic data since token construction may not closely follow parameters of natural speech or of any one talker. However, several hypotheses for the provision of extrinsic information from such data will be tested using one classification scheme based on extrinsic specification. The second group of theories assume that all information necessary for vowel recognition is *intrinsically* specified within the speech signal itself. This section will discuss and evaluate how well a number of these schemes are able to classify the perceptual data from Experiment I and the natural sppech data from Peterson and Barney (1952) as compared to the *SSB* target zones.

2.4.1 $F1$ x $F2$

An early approach to intrinsically specified vowel classification was utilized by Peterson and Barney in 1952. This approach consisted of plotting measurements from spectrograms of $F2$ against $F1$ for ten vowels from 76 speakers and hand-drawing ellipses around data points of like vowel quality. The plots utilized a frequency scale devised by Koenig (1949) which is linear to 1 kHz and logarithmic above. Although the number or percent of tokens correctly classified by this procedure was not reported in this study, Peterson and Barney do state that the ellipses enclosed approximately 90% of the data points in this manner. They go on to state that considerable overlap exists between /ER/ and /EH/, /ER/ and /UH/, /UH/ and /UW/, and /AA/ and /AO/, although the overlap between /ER/ and other vowels may be disambiguated by taking the low values of $F3$ found for /ER/ into consideration.

The outlines of the ten vowel ellipses (see Figure 2-14) from Figure 9 of the Peterson and Barney study were traced using a digital tablet and stored on a computer for use in a plotting program to generate the figures that follow. The plurality identifications for the 1674 synthetic tokens (rejected points not included) were then plotted, overlayed on the ellipses, in a $F2$-$F1$ space using the Koenig scale for the $F2$ axis. The results of these plots are shown in Figure 2-15 for all categories except /ER/ and Figures 2-16a-g for all tokens by z' plane.

Although there is considerable overlap, the data points seen in Figure 2-15 fall generally in the appropriate ellipses and reasonably good agreement is found between the /AE/, /AA/, and /UW/ data and their ellipses. The upper portions of the ellipses for the front vowels contain no data points, presumably due to the fact that these spaces originally enclosed the higher combinations of $F1$ and $F2$ for women and children in the Peterson and Barney study and such combinations were not within the constraints allowed for synthesis. The /EY/ and /OW/ identifications, categories not used in their study, generally share the ellipses of neighboring categories, /IH,EH/ and /AO,UW/ respectively, as might be expected. With $F3$ information not utilized, identifications from five non-retroflex vowel categories fall in the ellipse for /ER/.

Although no attempt was made to determine the percent correctly classified, classifica-

Figure 2-14: Vowel ellipses plotted in $F1$ x $F2$ space from Figure 8 of Peterson and Barney (1952).

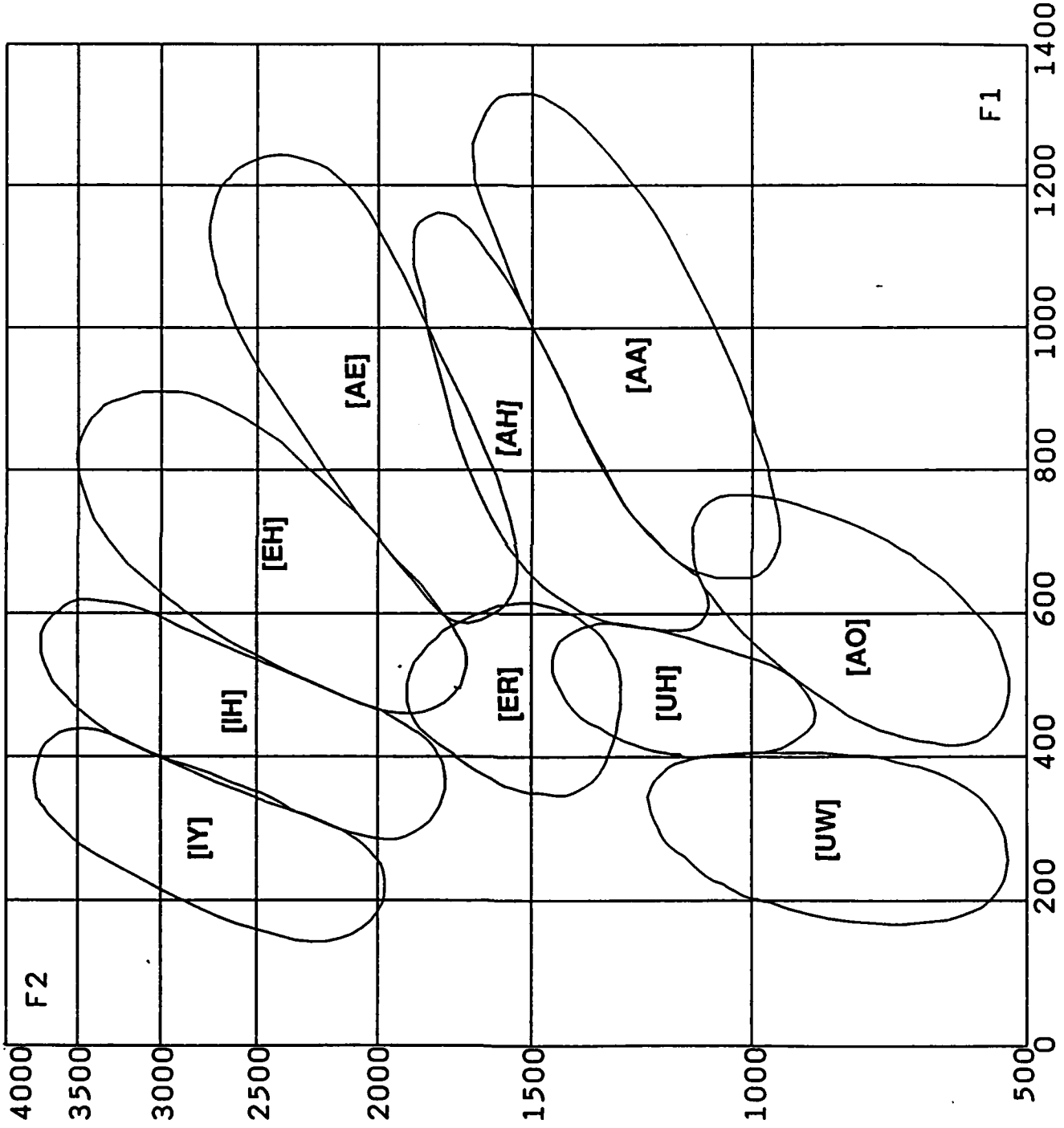


Figure 2-15: All plurality identifications with ellipses from Figure 2-14 in $F1$ x $F2$ space.

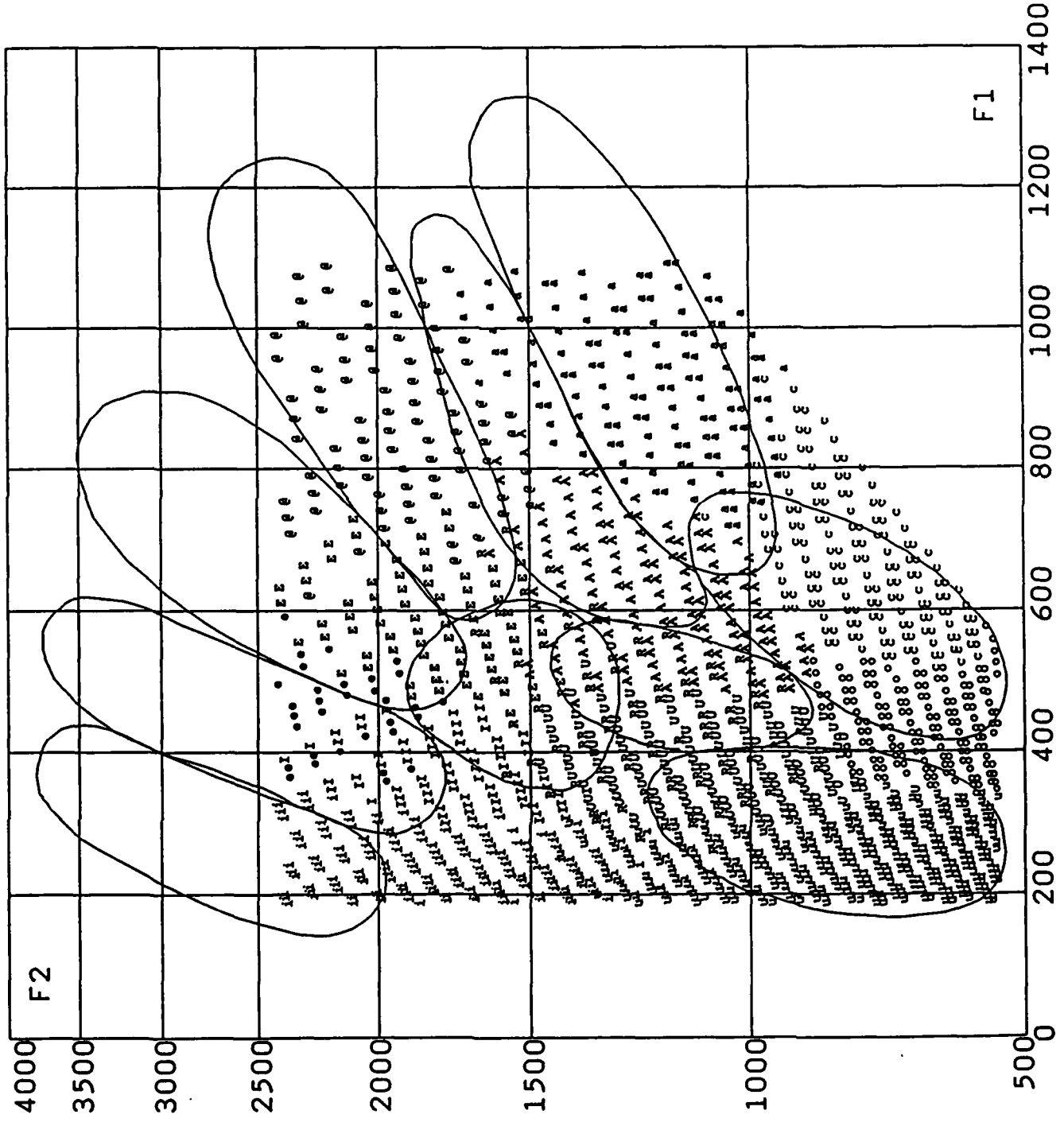


Figure 2-16: (a) Plurality identifications from the $z' = 0.80$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

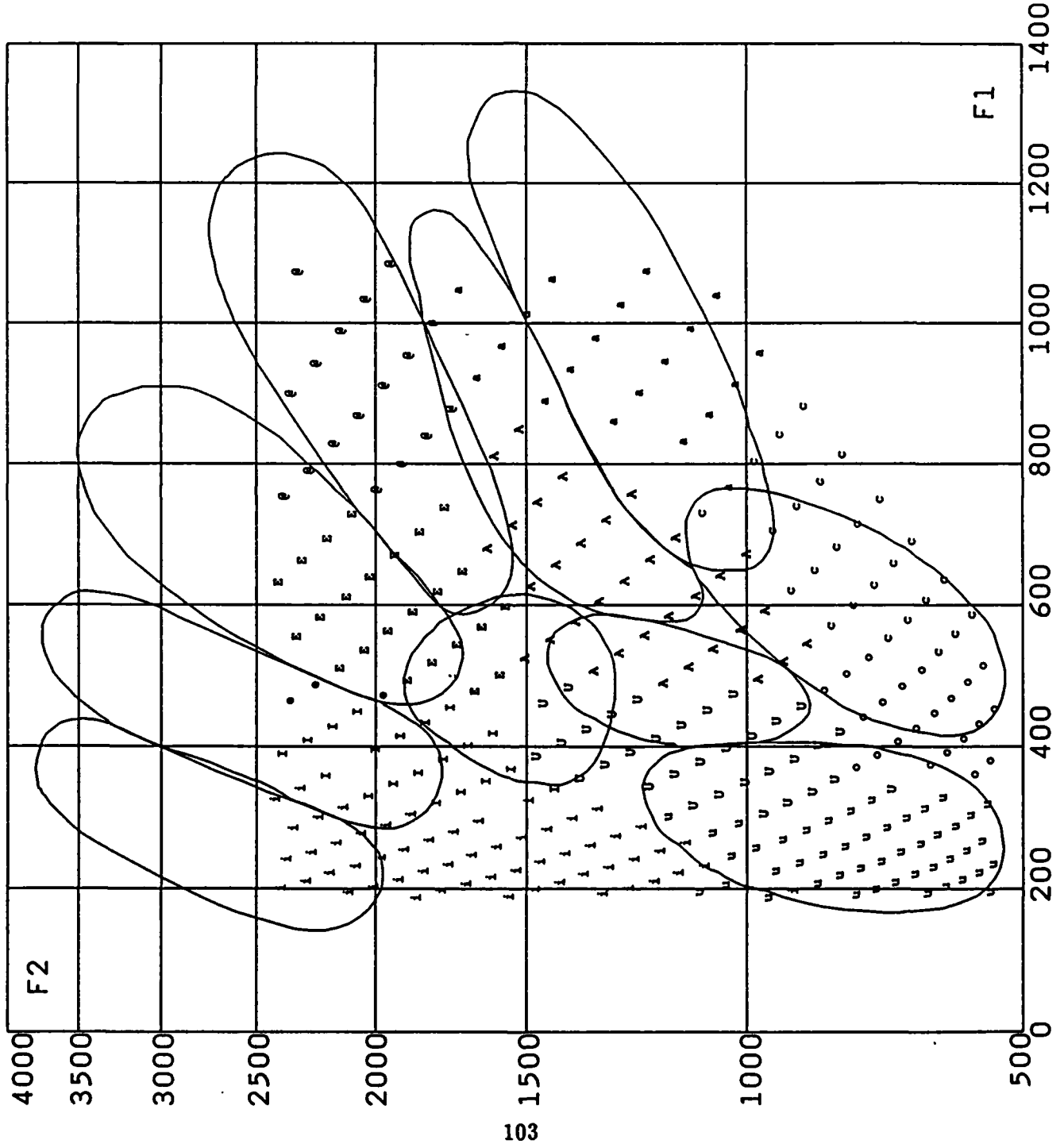


Figure 2-16: (b) Plurality identifications from the $z' = 0.75$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

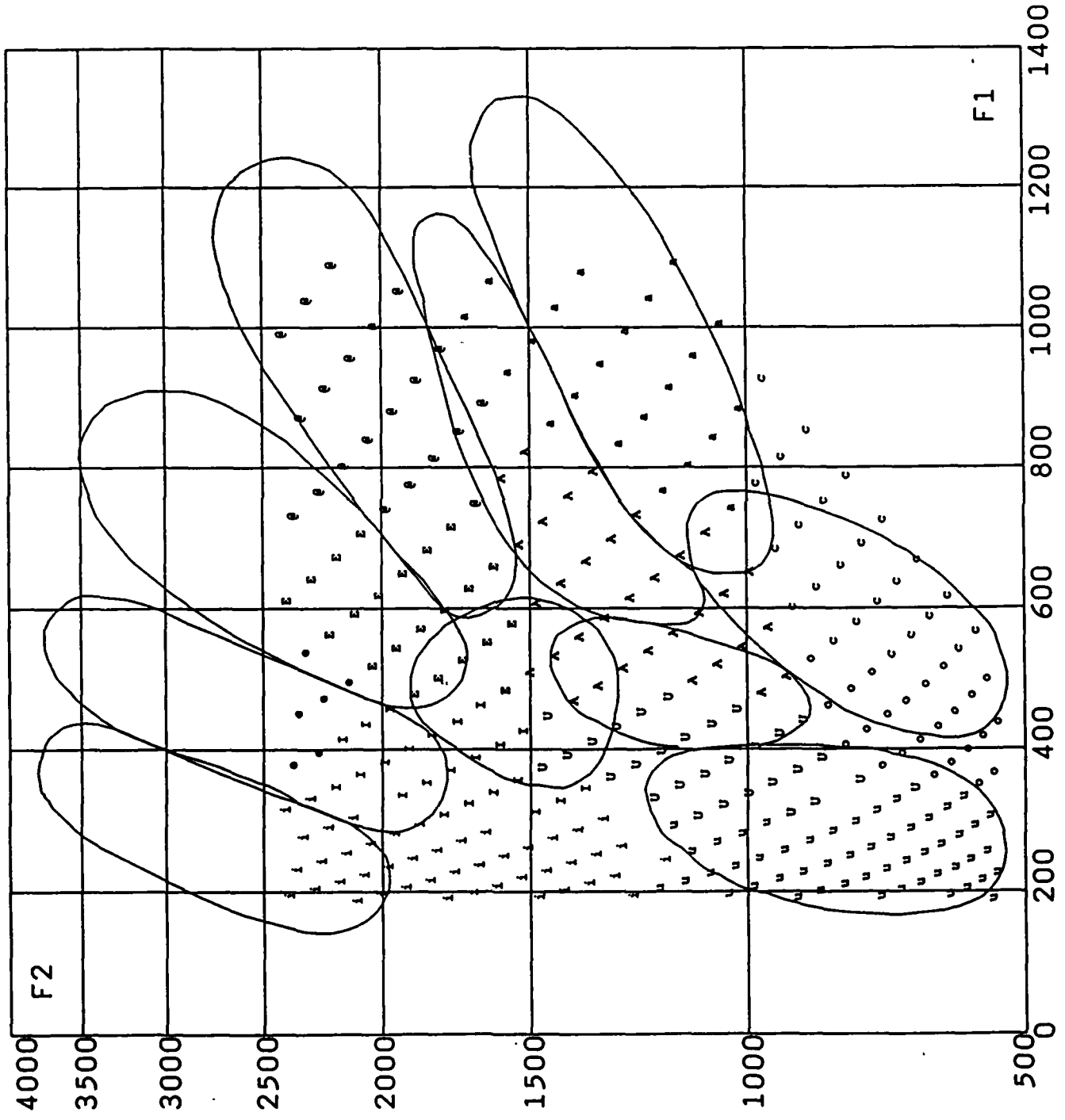


Figure 2-16: (c) Plurality identifications from the $z' = 0.70$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

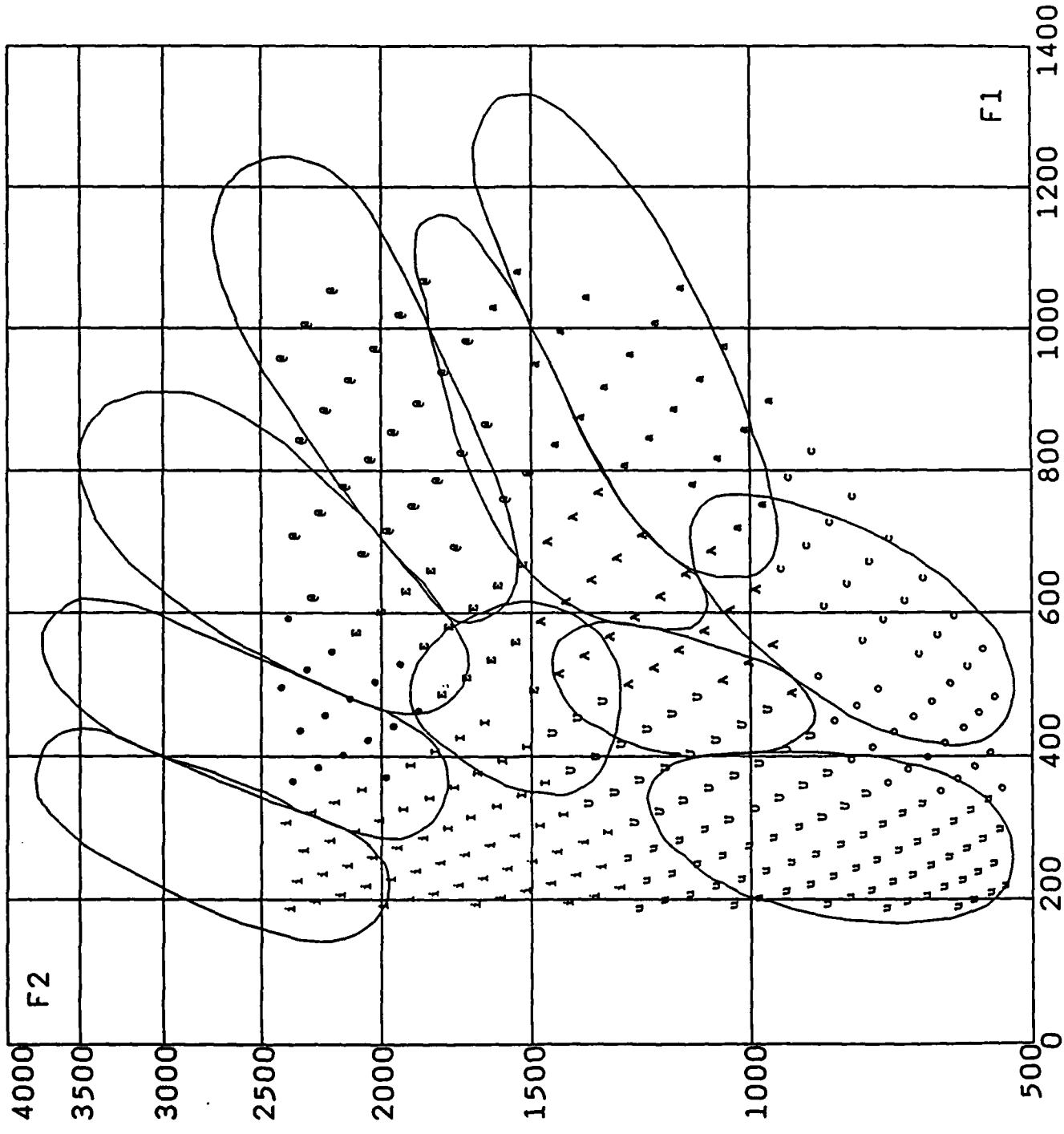


Figure 2-16: (d) Plurality identifications from the $z' = 0.65$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

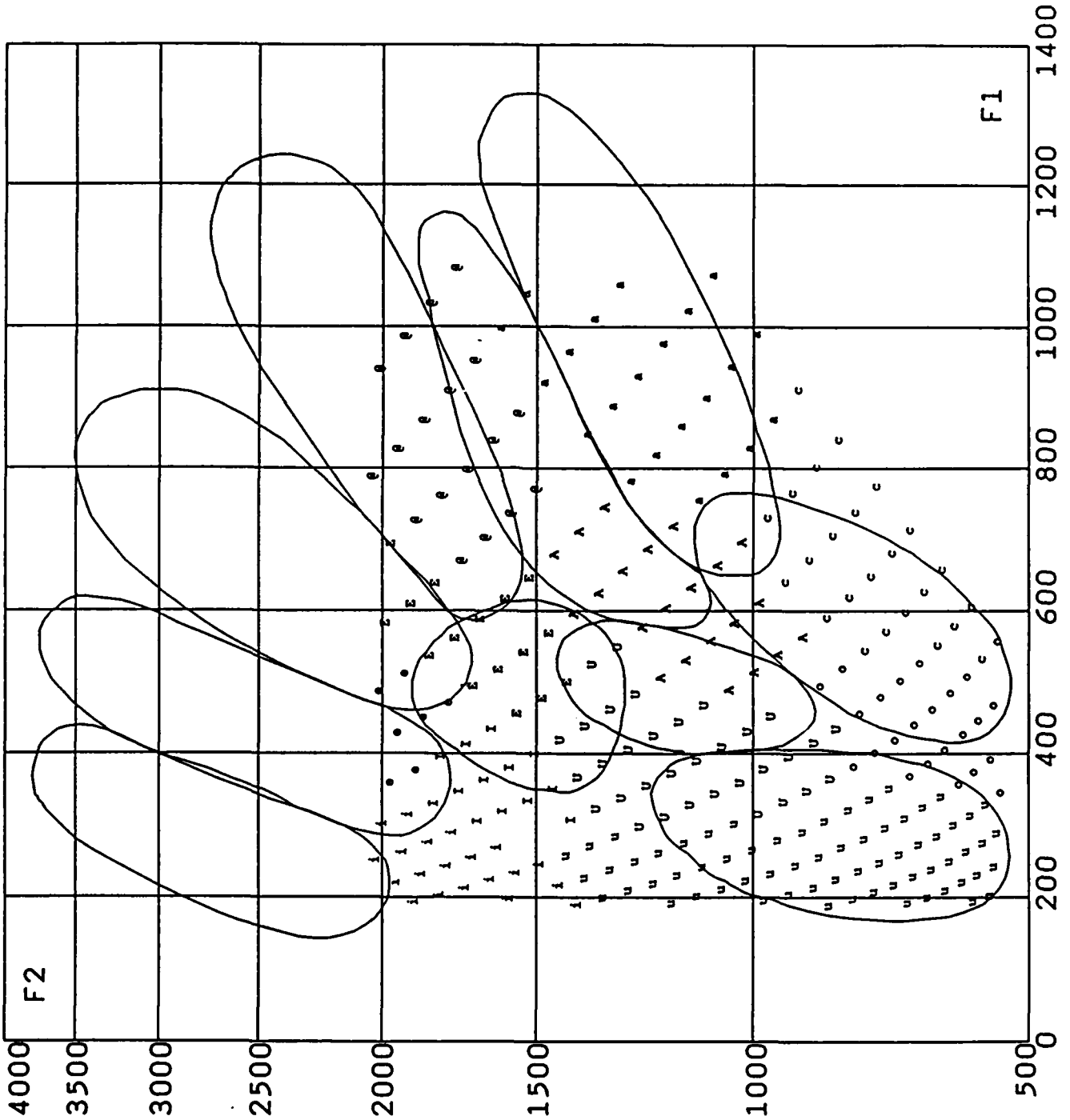


Figure 2-16: (e) Plurality identifications from the $z' = 0.60$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

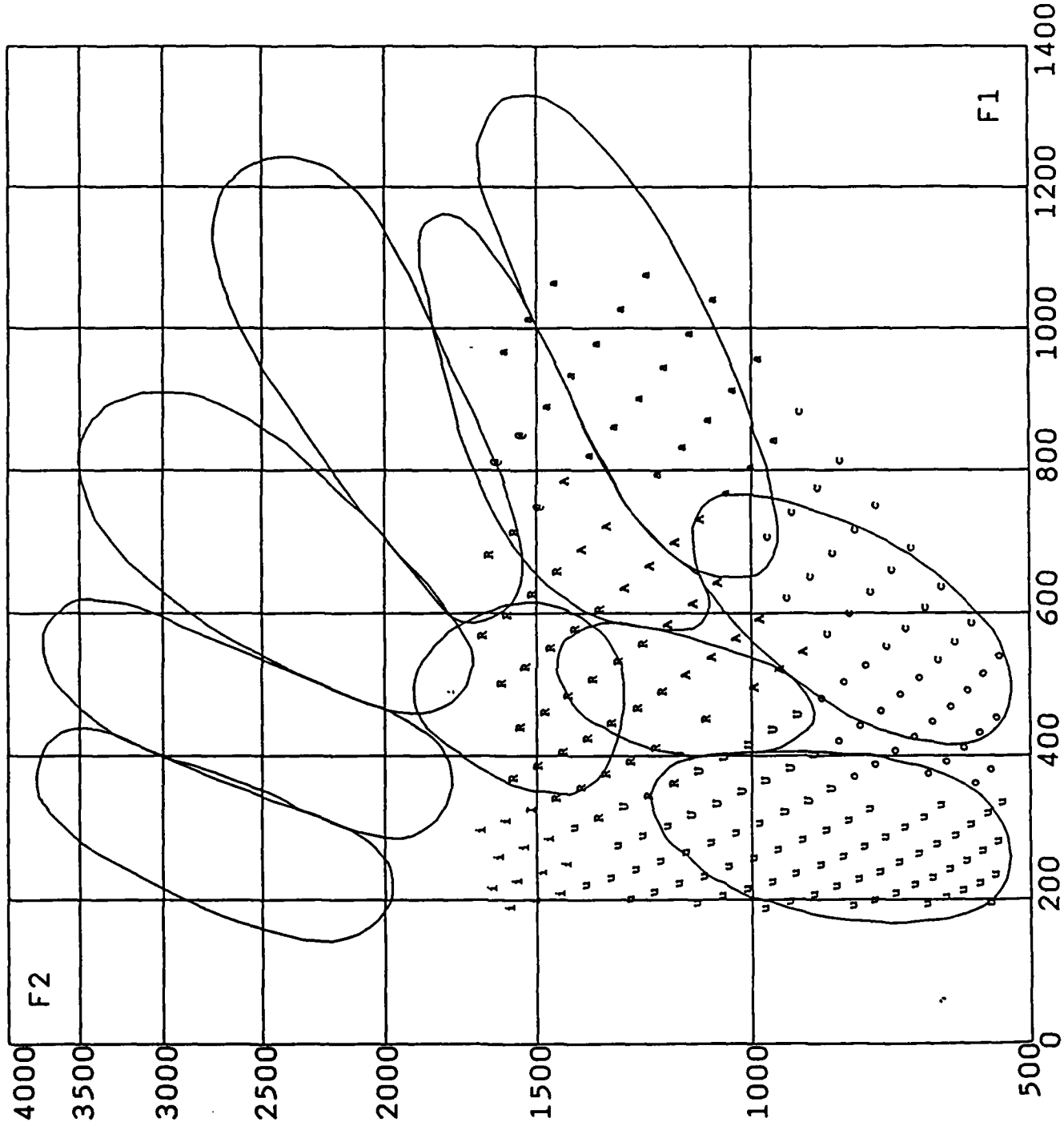


Figure 2-16: (f) Plurality identifications from the $z' = 0.55$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

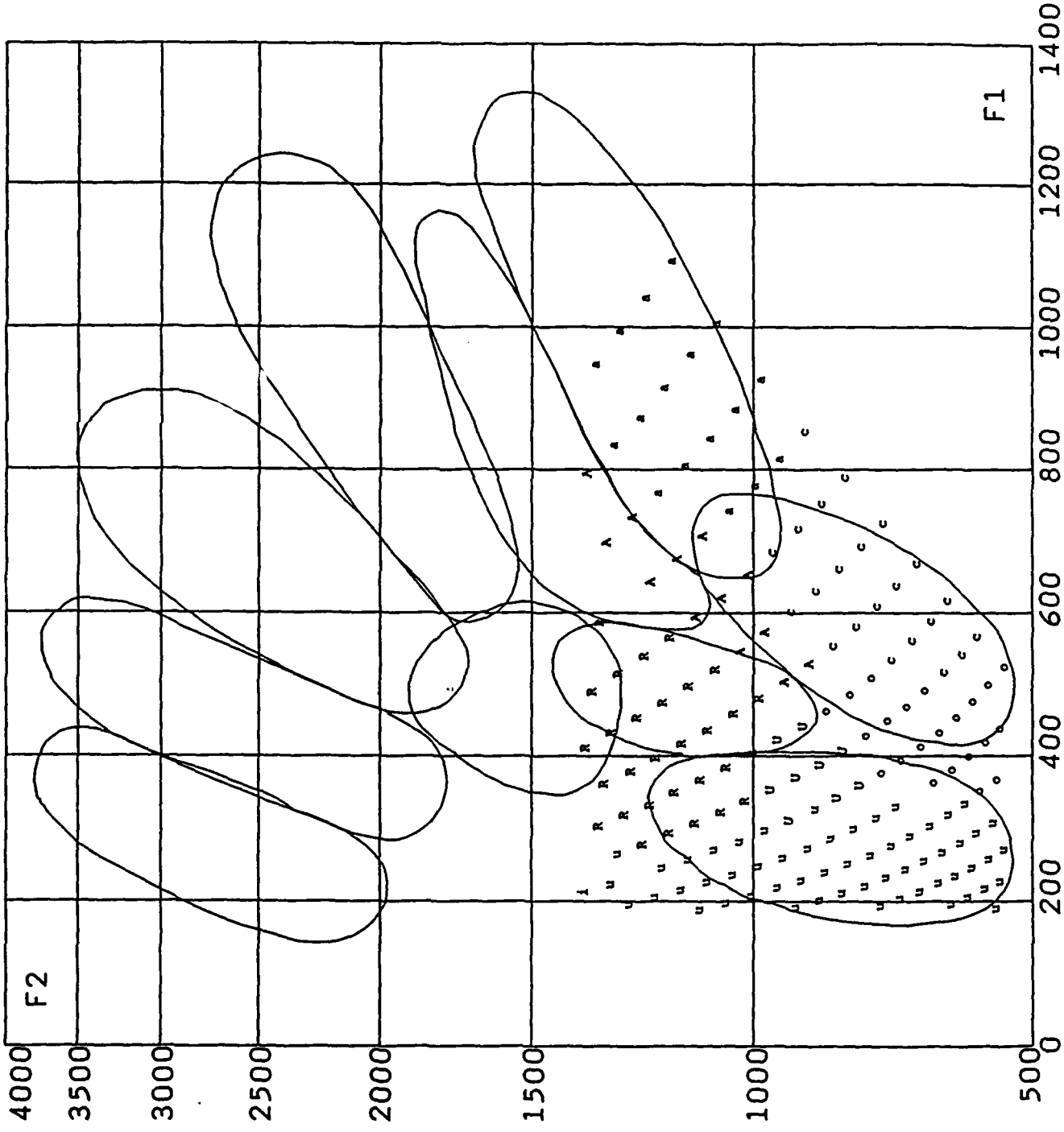
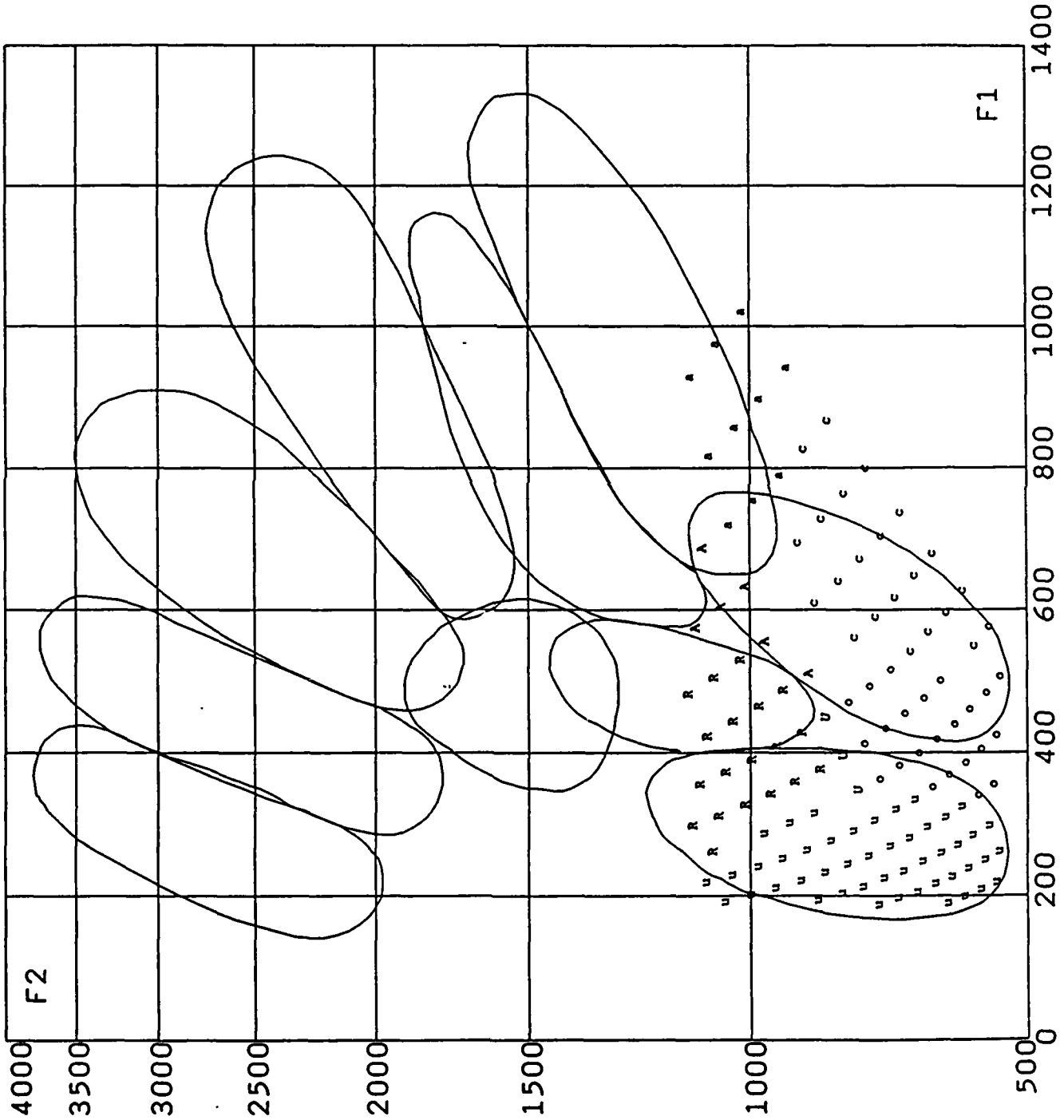


Figure 2-16: (g) Plurality identifications from the $z' = 0.50$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.



tion accuracy does not appear to significantly improve for even the primary planes, seen in Figures 2-16d-f, where tokens should most closely approximate natural speech. Considerable scattering of the data into ellipses adjacent to the appropriate categories is found for virtually every vowel group. Although new ellipses could be drawn based on the synthetic data with perhaps some improvement in classification accuracy, no simple delineation of vowel categories using this method is apparent that would yield high classification accuracy.

2.4.2 Comparison of synthetic and natural speech-based target zones

The vowel classification accuracy using the natural speech-based (*NSB*) target zones and the synthetic speech-based (*SSB*) target zones will now be estimated and described for several different data sets. These classifications are made possible by a computer program which creates three-dimensional digital maps of the zones using *APS* coordinates and then references these maps for the locations of labeled vowel tokens. The mapping resolution employed was 0.001 log units for both the *NSB* and *SSB* zones. Data points falling on zone boundary lines are counted as belonging to that zone. Since the shapes of *SSB* zones change at 0.05 log unit intervals in the z' dimension, the way in which the classification program deals with data points lying between the specified intervals is of importance. The program does not interpolate the zone boundaries between specified z' planes, but rather, utilizes the specifications nearest the data point of consideration. Thus a data point falling less than 0.025 log units behind a specified z' plane would be classified with the zones mapped from the z' plane in front of it. The *NSB* map boundaries reflect those shown in Figure 1-9 from Chapter 1. Map areas between boundary lines or outside zones are considered unclaimed space.

The preliminary results of classification are shown in Table 2.10 for the *NSB* target zones and for the *SSB* target zones. The data set "ALL" contains all 27,600 responses from the 16 subject response sets. The "PLURALITY" data set contains the plurality identifications excluding the rejected points. The "AGREE" data set contains the 320 tokens having unamimous identification agreement. For each of these data sets, there is an additional set labeled "(primary)" which contains the same data limited to the primary planes. Additionally, the "CID" data set contains 599 points representing measurements

Table 2.10: Preliminary vowel classification using *NSB* and *SSB* target zones.

Data Set	N total	*NSB* zones # corr.	*NSB* zones % corr.	*SSB* zones # corr.	*SSB* zones % corr.
ALL	27600	5237	19.0	20424	74.0
ALL (primary)	14112	4165	29.5	10623	75.3
PLURALITY	1674	354	21.1	1674	100.0
PLURALITY (primary)	862	278	32.3	862	100.0
AGREE	320	76	23.8	320	100.0
AGREE (primary)	170	62	36.5	170	100.0
CID	599	597	99.7	481	80.3
P & B	1520	1378	90.7	960	63.2

from natural speech samples and data from the literature, and the "P&B" set, 1520 points representing the natural speech measurements made by Peterson and Barney (1952) of 10 vowels in [hVd] context spoken twice by 33 men, 28 women, and 15 children. All data points are expressed as x', y', and z' coordinates.

The classification accuracy for the *NSB* target zones reflected by the percent correct scores from column 4 of Table 2.10 is relatively poor for all synthetic speech data sets compared to the scores for the *SSB* zones in column 6. However, the classification accuracy regarding these sets is underestimated, due to two factors. The first factor is that certain identification categories (/OW,EY/) used in the synthetic data sets are not currently represented by *NSB* target zones, and thus could not possibly be correctly classified. The number of identifications in the /OW/ and /EY/ categories for each data set are shown in column 3 of Table 2.11. When these tokens are subtracted from the totals, the resulting percentages of correct classification increase somewhat. The second factor concerns the difference between the space occupied by the *NSB* target zones and the space utilized for synthesis. Column 4 from Table 2.11 indicates the number of tokens falling in unclaimed space, i.e. areas between or outside of target zones. Detailed examination of the classifica-

Table 2.11: Corrected vowel classification using *NSB* and *SSB* target zones.

Data Set	N total	*NSB* zones # EY,OW	# UCS	corrected N	% corr.	*SSB* zones # UCS	corrected N	% corr.
ALL	27600	2902	13321	11377	46.0	752	26848	76.1
ALL (primary)	14112	1610	5720	6682	61.4	288	13824	76.8
PLURALITY	1674	171	804	699	50.6	—	—	100.0
PLURALITY (primary)	862	100	343	419	66.3	—	—	100.0
AGREE	320	10	198	112	67.9	—	—	100.0
AGREE (primary)	170	10	95	65	95.4	—	—	100.0
CID	599	—	1	598	99.8	10	589	81.7
P & B	1520	—	62	1458	94.5	71	1449	66.3

tion data indicates that less than 10 tokens fall between narrow boundary lines, thus the majority of unclaimed tokens fall outside the *NSB* target zones. If the number of unclaimed tokens along with the /OW/ and /EY/ identifications are subtracted from the totals of each data set, the percentages of correct classification increase considerably, although over half the data in each synthetic data set has now been excluded from consideration. Note that overall, the classification accuracy of the *NSB* zones for the synthetic data increases with increasing data set identification agreement.

The number of tokens falling in unclaimed space are also used to reduce the total number of tokens under consideration for the *SSB* zones. However, now the majority of tokens in unclaimed space represent only locations between zones, not outside of zones. This could provide the advantage of classifying these tokens with multiple identifications, rather than just listing them as unclaimed, although this will not be attempted at this time. The percent correct classification by the *SSB* zones of all identifications collected from Experiment I after correction for unclaimed space is 76.1%. This percentage should represent the maximum possible correct for classifying this data set, since the *SSB* zones were constructed to classify the plurality identifications from this data correctly, and the

remaining 23.9% of the identifications must represent the minimum number of differences from the plurality identifications. This implies that the percentages correct for the "ALL" data sets using the *NSB* target zones may actually reflect an even higher accuracy than indicated.

The data sets limited to identifications in the primary planes were included to evaluate whether or not the identifications of tokens located in the z' planes of the *APS* most associated with natural speech are classified with higher accuracy than identifications from all z' planes. For the *SSB* zones, only a very slight increase in accuracy is found when all identifications are limited to the primary planes. This reflects the results discussed in Section 2.3.10, where the average pair-wise agreement by z' plane remains relatively constant. However, for the *NSB* zones, an average increase in classification accuracy of over 15% is found for each data category when the data set is limited to the primary planes. This suggests that the *NSB* zones are more accurate at classifying identifications of synthetic tokens located in *APS* areas shared by natural speech, although the difference in accuracy may predominantly reflect the inability of the *NSB* zones to capture differences in the z' dimension outside the range of natural speech.

The classification accuracy for the natural speech data sets (CID and P & B) is somewhat reversed from that of the synthetic data sets. The *NSB* target zones appear to be considerably more accurate at classifying natural speech than are the *SSB* zones. The higher accuracy for the *NSB* zones could be anticipated, given the consideration that these zones were constructed using these data sets as their basis. However, the classification accuracy of the *SSB* zones is lower than anticipated and requires at least speculative explanation. Since the *SSB* zones were developed utilizing tokens approximating a male voice, the possibility that the female and child data may be the source of the poor classification accuracy was investigated. Separate classifications were run for the tokens from men, women, and children of the data set from Peterson and Barney (1952) with the *SSB* zones. The results indicated that the male data was classified only slightly more accurately (68.5%) than the overall perecentage and that the data from women and children slightly less accurately, 65.2% and 63.2% respectively. These differences in classification accuracy cannot be considered accountable for the low overall accuracy.

To further investigate the lower than expected classification accuracy of the *SSB* zones for natural speech data, the errors in classification by the *SSB* zones were more closely examined. This examination yielded two major sources of errors. The first error source was the erroneous classification of tokens as /EY/ or /OW/. The /EY/ classification errors accounted for 81.6% and 69.3% of the errors made on /EH/ and /IH/ identifications respectively, and the /OW/ errors accounted for 60.9% and 34.2% of the errors made on /AO/ and /UH/ respectively. Overall, tokens classified as /EY/ or /OW/ accounted for 41.7% of the total errors. If these errors along with tokens classified in unclaimed space are subtracted from the total possible, classification accuracy increases to 77.1%. These results suggest that tokens located in at least portions of the zones for /EY/ and /OW/ may be identified as vowels in neighboring zones when /EY/ and /OW/ are not among the possible responses for identification. Remapping these areas without /EY/ and /OW/ among the possible identification responses may clarify this issue and increase the classification accuracy of the *SSB* zones.

A second source of error gives consideration to the possibility that the resolution of the zones is too coarse to accurately estimate the true boundaries between zones. That is, the step size utilized in Experiment I for sampling the *APS* is too large to adequately reflect boundary information. The misclassifications for all vowels except /ER/ are shown as their intended identifications by grouping along the nerest z' axis in Figures 2-17a-c along with the zones appropriate for that z' region. Visual inspection of the misclassified tokens in *APS* indicates that the majority of errors not located in the zones for /EY/ or /OW/ occur at or near boundaries between zones. Examination of the plurality frequencies and ratings sums (See Figures 2-10a-g and 2-11a-g) along these boundaries suggest that some of these boundaries are weak, reflected by ties or low plurality and rating values, and that these boundaries may shift given a higher resolution mapping of these areas. A more complete appraisal of the classification accuracy of natural speech data with zones based on synthetic speech identifications should then perhaps rest with further research investigating boundary areas between zones in finer detail. The next section will discuss a preliminary attempt toward this goal.

Figure 2-17: (a) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.75$ plane which were misclassified by the *SSB* target zones.

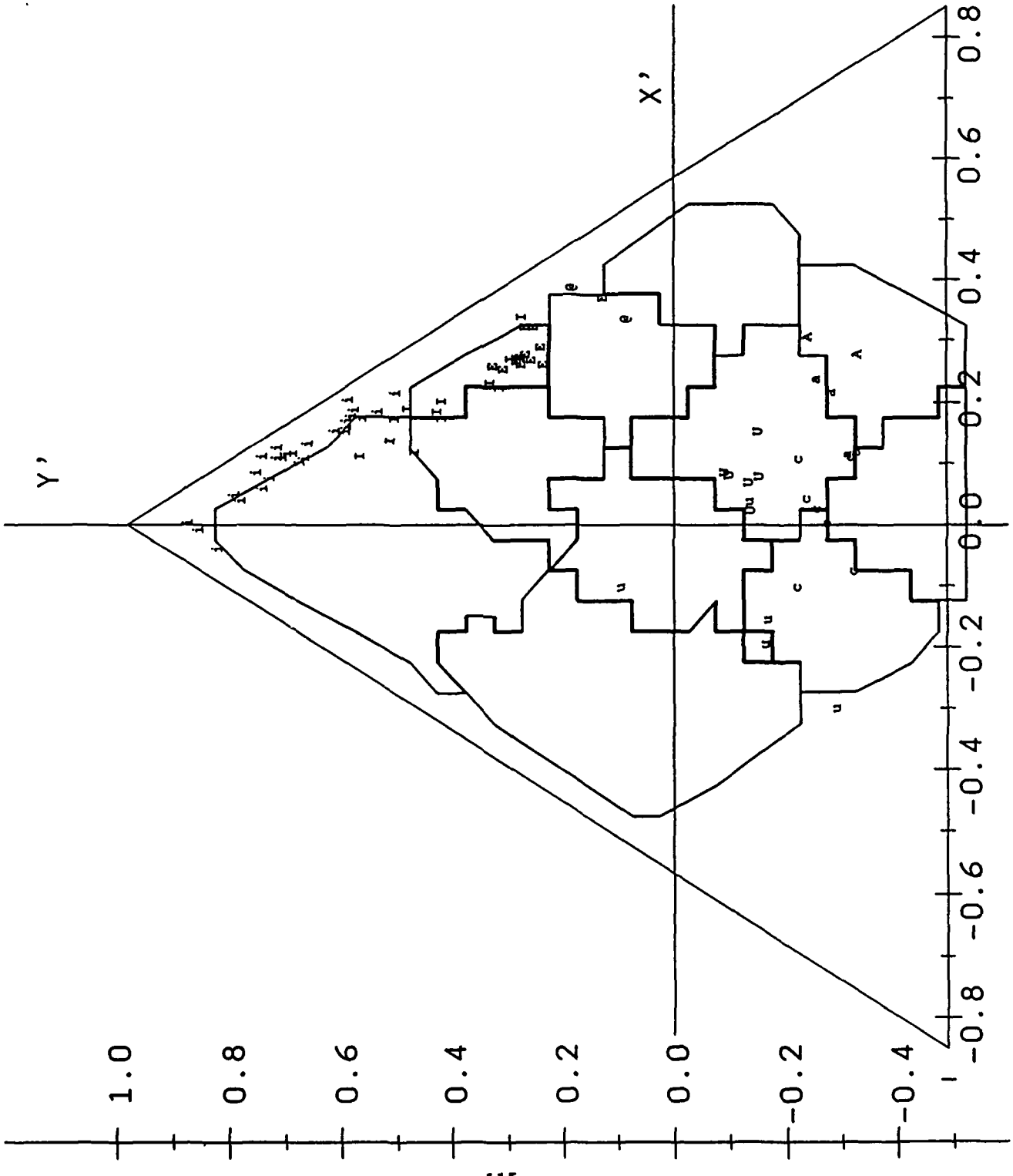


Figure 2-17: (b) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.70$ plane which were misclassified by the *SSB* target zones.

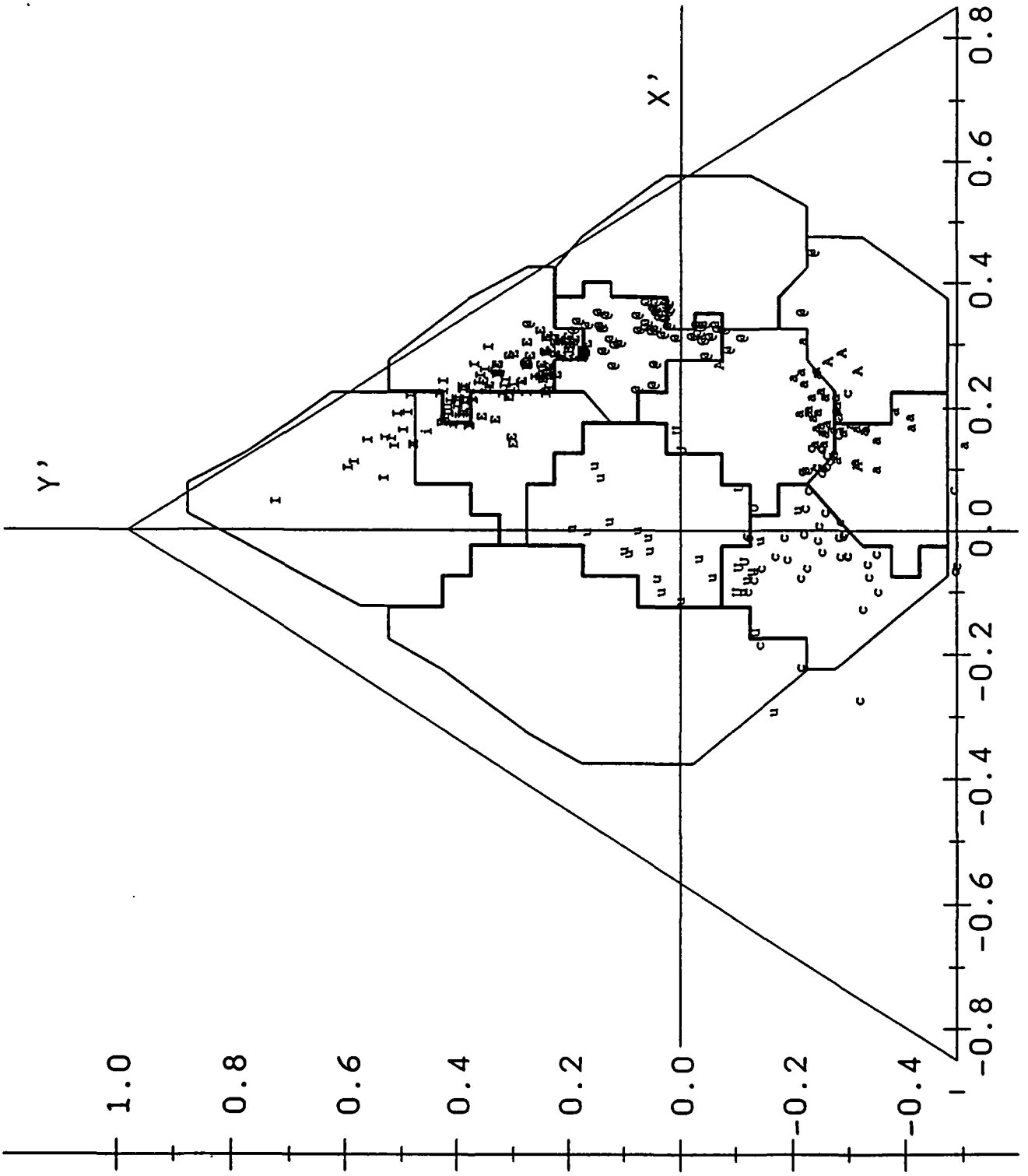
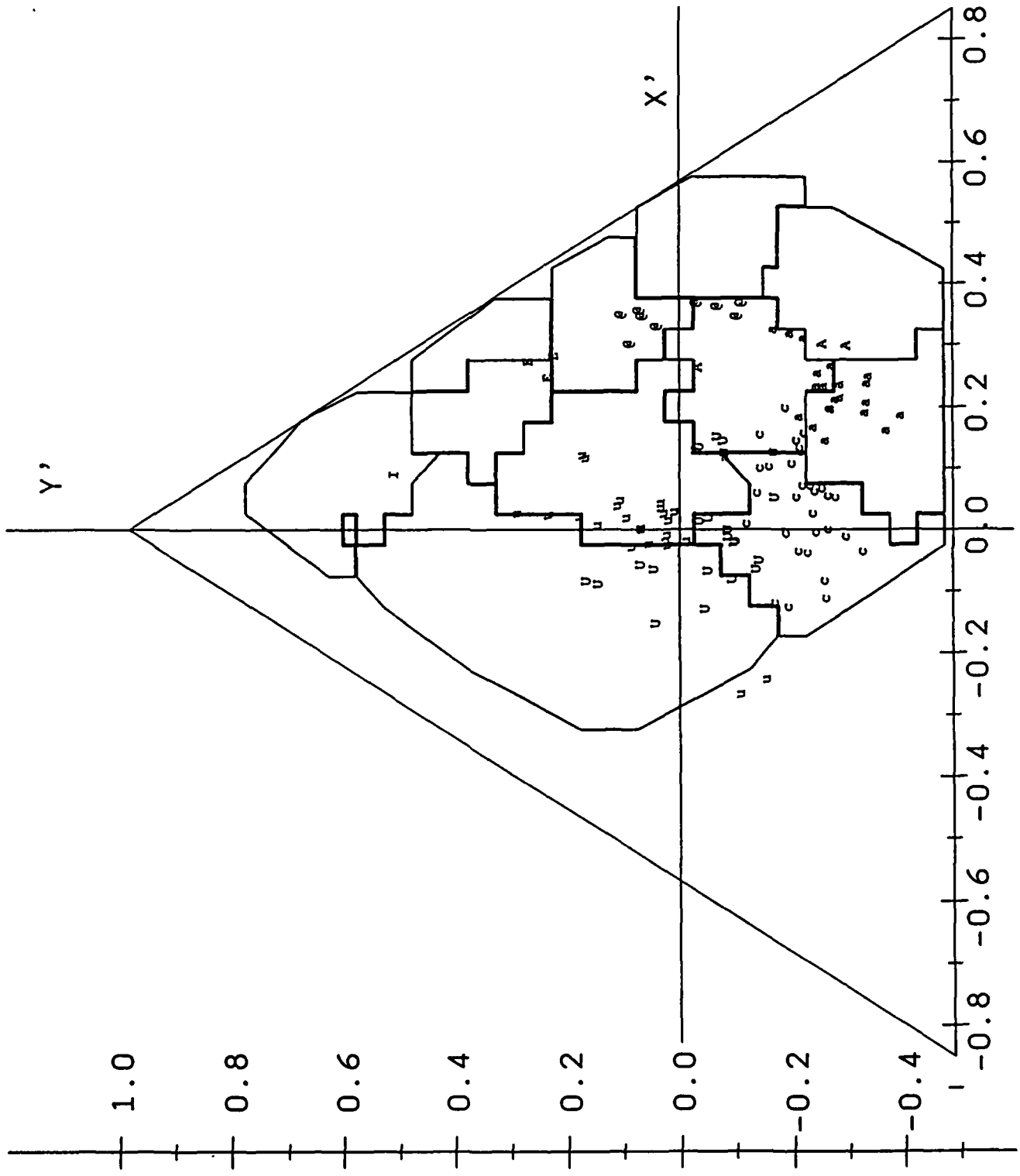


Figure 2-17: (c) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.65$ plane which were misclassified by the *SSB* target zones.



Classification using high-resolution target zones

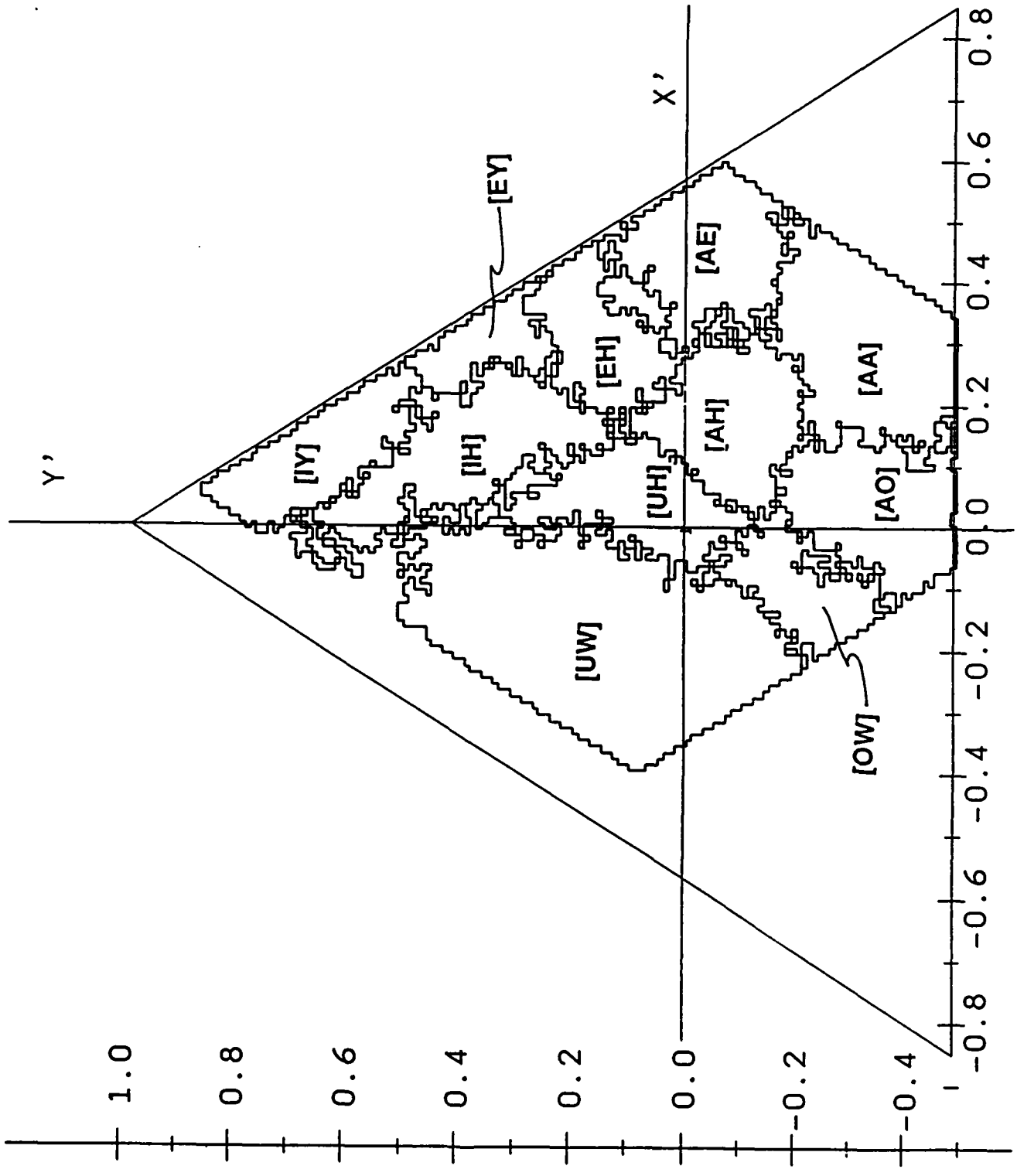
Miller and Hawks (1989) presented the results of a *APS* mapping experiment identical methodologically to the experiment presented in this paper with the exception of two differences. The synthesized tokens represented equidistant .01 log unit points in one z' plane ($z' = 0.70$) of the *APS* and were identified by only one subject highly trained in phonetics. This experiment yielded identifications to 7703 synthetic vowels for mapping this single z' plane at 5 times the resolution of the main experiment presented here.

Target zones, shown in Figure 2-18, based on these identifications were constructed automatically by a computer algorithm which encloses all points of like identification which are adjacent to one another. These zones were then used to classify the "PLURALITY" data limited to the $z' = 0.70$ plane and the "P&B" data limited to points located within the $z' = 0.70$ plane (± 0.49 log units). The results of classification with these zones (referred to as *HiR SSB*) along with similar results for the *NSB* and *SSB* zones are shown in Table 2.12. Note that classification accuracy for the "P&B" natural data set with the *HiR SSB* zones is over 12% higher than the lower resolution *SSB* zones. This result suggests that mapping at least boundary areas in *APS* at higher resolution may indeed increase classification accuracy of natural data. Additionally, the utilization of highly-trained subjects and equivalent response sets for identification of both synthetic and natural data may also aid in more accurate classification.

Table 2.12: Classification using *NSB*, *SSB*, and *HiR SSB* target zones.

Data Set	Zone Type	N total	# corr.	# UCS	% corr.
PLURALITY	*NSB*	296	105	-159	76.6
	SSB	296	296	—	100.0
	HiR SSB	296	224	-10	78.0
P & B	*NSB*	843	757	-30	93.1
	SSB	843	530	-18	64.2
	HiR SSB	843	629	-21	76.5

Figure 2-18: Location in APS $x'y'$ coordinates of computer-constructed high-resolution target zones based on data from Miller and Hawks, 1989.



2.4.3 Classification using bark differences

Syrdal and Gopal (1986) present an intrinsic vowel classification scheme which preserves articulatory feature information, and thus specifies vowel quality. These researchers found that, when $F0$, $F1$, $F2$, and $F3$ were transformed to critical band (bark) values, $F1-F0$, $F2-F1$, $F3-F2$ differences delineated the features high/non-high, compact/noncompact, and front/back respectively. The point of delineation for the binary feature distinction was 3 bark with differences less than 3 bark represented as a "+" for that feature and differences equal to or greater than 3 bark a "−". The resulting binary feature system for American English vowels[5] are shown in Table 2.13. Syrdal and Gopal state that the bark differences for $F1-F0$ corresponding to vowel height straddle this boundary for the vowels /AO/ and /ER/, and thus cannot be classified along this dimension, although both generally exceed the 3 bark criterion. The vowel /ER/ however can be solely classified on the basis of an additional spectral distance measure of $F4-F3$, should this information be available. The bark difference classification scheme was implemented on computer and applied to the synthetic data sets used in section 2.4.2. Values of $F0$, $F1$, $F2$, and $F3$ for the data were first transformed to bark values following the critical band scale approximation of Zwicker and Terhardt (1980). This approximation was modified with Traunmüller's (1981) low-frequency end correction for frequencies below 250 Hz. The $F1$–$F0$, $F2$–$F1$, and $F3$–$F2$ bark differences were then calculated and tested with the vowel feature classifications from Table 2.13. The results of these classification analyses are shown in Table 2.14. In comparing the percentages correct for the bark difference metric for the synthetic data sets with the results of the *NSB* zones in Table 2.11, we find that the classification accuracy of the bark difference metric is considerably higher for all synthetic data sets except the agreements limited to the primary planes, for which it is about comparable. The classification accuracy for the natural data set from Peterson and Barney (1952) is reasonably good, and improves further with the elimination of the /ER/ identifications to 89.0%. However, the classification accuracy of the bark difference metric is inflated due to limitations of the vowel features utilized in the classification. Inspection of Table 2.13 reveals that the feature specifications listed are not

[5]The features shown in parentheses were not included in Syrdal and Gopal (1986) and have been determined based on the features best fitting identifications in the synthetic data sets.

Table 2.13: Vowel feature system using bark-difference dimensions from Syrdal and Gopal (1986). Features in parentheses are based on best fit to synthetic data.

Vowel	$F1 - F0$ < 3 bark	$F2 - F1$ < 3 bark	$F3 - F2$ < 3 bark
IY	+	−	+
IH	+	−	+
EY	(+)	(−)	(+)
ER	(+)	−	+
EH	−	−	+
AE	−	−	+
AH	−	−	−
AA	−	+	−
AO	(−)	+	−
OW	(+)	(+)	(−)
UH	+	−	−
UW	+	−	−

Table 2.14: Vowel classification using Syrdal and Gopal (1986) classification scheme.

Data Set	# corr.	N total	% corr.
ALL	16857	27600	61.1
ALL (primary)	10276	14112	72.8
PLURALITY	1094	1674	65.4
PLURALITY (primary)	673	862	78.1
AGREE	255	320	79.7
AGREE (primary)	160	170	94.1
P & B	1291	1520	84.9

capable of distinguishing all twelve vowel categories, but rather, six vowel category groups. The ambiguities include the tense/lax pairs /IY-IH/, /EH-AE/, /AA-AO/, and /UH-UW/, as well as /ER/ and /EY/, which are indistinguishable from /IY/ and /IH/. Syrdal and Gopal (1986) state that this classification scheme cannot classify vowels properly without consideration of additional temporal parameters not captured by the bark difference metric. Since the synthetic tokens data classified here were constructed with identical duration parameters, this dimension cannot be utilized as a classification parameter. Thus, Table 2.14 actually reflects the classification accuracy of six categories of feature groupings and not 12 vowel categories. An accurate classification by vowel category utlizing this classification metric would yield correct classifications for only two categories /AH/ and /OW/, since all other categories are confusable. While such a metric demonstrates high sensitivity to vowel features, its ability to classify vowels by traditional categories is limited without the inclusion of additional information. This limitation suggests that this metric may not be appropriate for the classification of certain synthetic data.

2.4.4 Vowel classification utilizing extrinsic specification

As was mentioned previously, a number of vowel classification theories rely on information distributed across all vowels of a talker for normalizing or providing a reference framework for vowel classification. Often the information required is traditional acoustic parameters, i.e. formants, but some extrinsic theories also make use of formant bandwidths or formant amplitudes[6]. Collecting information from a reference series of vowels attributed to a single talker poses problems for classifying identifications for synthetic data as in Experiment I. This is because the motivation of the experiment was to uniformly map all the possible vowel space appropriate for a male talker and does not provide a set of reference vowels, per se. However, classification of the data sets used previously will be attempted with one such classification scheme utilizing two approaches for establishing a vowel reference framework for extrinsic specification. The classification scheme to be tested is from Neary (1977) and was selected because of its superior normalization performance in a comparative study by Disner, 1980. This scheme incorporates several different procedures, two of which will be

[6]For more detailed discussions of extrinsic classification schemes, see Disner, 1980 and Neary, 1989.

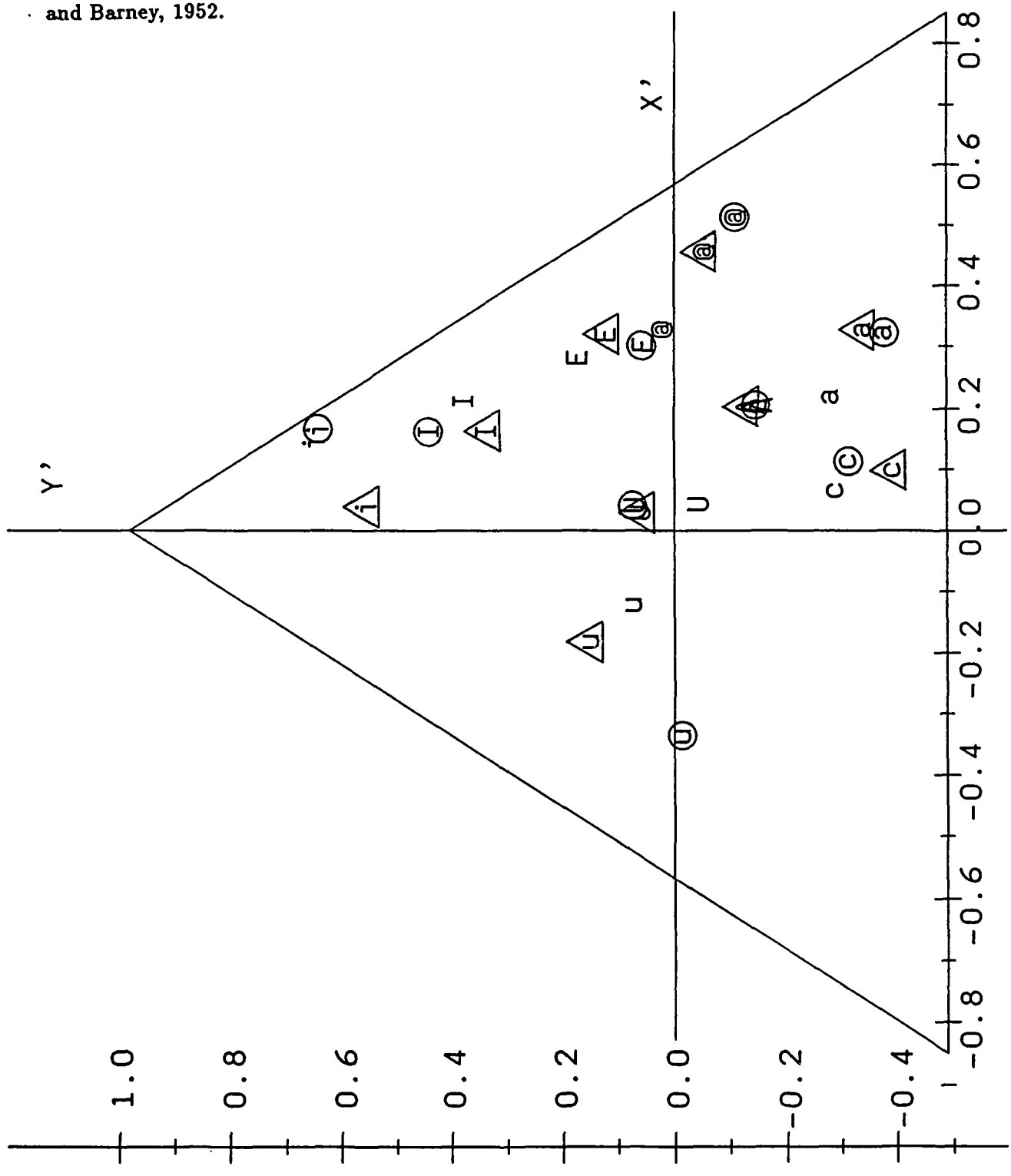
utilized here. Both procedures first transform formant values to their natural logarithms, notated $G1$, $G2$, and $G3$. The *CLIH* (constant log interval hypothesis) procedure then normalizes vowels by subtracting the average of the mean values of $G1$ and $G2$ for all vowels of a given talker from each $G1$ and $G2$. The *CLIH3* procedure is similar except that three independent parameters are established. Here the average value of $G1$ is subtracted from all values of $G1$, the average value of $G2$ is subtracted from all values of $G2$, and likewise for $G3$.

Two approaches for establishing a reference framework for extrinsic specification were utilized for classification. Both approaches assumed that all vowel tokens in Experiment I could be considered those of a single talker. The first approach hypothesized that the reference framework might be best represented by an exemplary token from each vowel category used in the mapping experiment. These exemplary tokens were determined by first selecting the token with the highest identification agreement for each vowel category from Experiment I. If several tokens were tied by this criterion, the token receiving the highest sum of ratings was selected. The averages of the natural log values of $F1$, $F2$, and $F3$ utilized in the synthesis specifications of these exemplary tokens served as the normalizing parameters for classification using this, the *EX* reference framework.

The second approach hypothesized that the reference framework might be best represented by the average location of all tokens identified as belonging to each vowel category used in the mapping experiment. The average x', y', and z' value was found for all tokens associated with each vowel category based on the plurality agreements. Formant values from these averages were calculated and the averages of their natural log values served as the normalizing parameters for this, the *AV* reference framwork.

The locations in *APS* of the token sets making up the reference frameworks for the two approaches are shown in Figure 2-19 along with the locations of the male averages from Peterson and Barney (1952). The vowel symbols in circles represent the first approach exemplary tokens, the symbols in triangles represent the second approach average tokens, and the plain symbols are the Peterson and Barney averages. Since the location of these points vary along the z' axis, manipulations were made to equalize these differences to a unitary z' plane. The locations of the Peterson and Barney points have been calculated

Figure 2-19: Locations in APS $x'y'$ coordinates of EXEMPLARY (circles) and AVERAGE (triangles) vowel reference frameworks used for extrinsic specification in Neary-type classification procedure along with locations (for comparison) of average male vowels from Peterson and Barney, 1952.



assuming a fixed $F0$ of 132 Hz, the value used for the synthetic tokens, instead of the actual average values. The values of $F3$ for all synthetic tokens and the Peterson and Barney data have been arbitrarily set to 2528 Hz, yielding a z' of .700 for all data points. This manipulation in effect allows the figure to reflect only $F1$ and $F2$ information in the relative distances between points.

Note in this figure that the tokens for the "point" or "corner" vowels, [IY,AE,AA,UW], representing extreme points of vowel articulation, are distanced furthest apart for the exemplary tokens than for the averaged tokens, and are closest together for the natural data tokens. Tokens representing vowels interior to the points of extreme articulation are generally more closely grouped across the three sets of tokens with a unitary location for the points representing the centralized vowel /AH/. These distance differences suggest that listeners may tend to separate the exemplary vowels at the extreme points of articulation, further apart than would be generally found in natural speech, in order to accomodate the vowel space being sampled. This follows the principle of maximum perceptual contrast, as demonstrated by Liljencrants and Lindblom (1972), whereby phonetic categories of vowel systems are distributed within the acoustic vowel space so as to maximize the distance between categories.

A similar expansion of vowel space can be seen in the data from Picheny, Durlach, and Braida (1986). In this study, the vowels [IY,AE,AA,UW] represented in a $F1$ by $F2$ space reach more extreme points of articulation when talkers are instructed to speak as clearly as possible, compared to the same vowels when spoken in a conversational manner.

The extrinsic normalizing parameters for classification of the natural data set from Peterson and Barney (1952) were established for each individual talker by calculating the average $G1$, $G2$, and $G3$ from the two vowel sets of each of the 76 talkers. The vowel category /ER/ was not included in the normalizing parameters for the synthetic or the natural data sets and was therefore also excluded as a category from the classification procedure. This exclusion reduced the total number of data points for consideration from 1674 to 1592 for the PLURALITY synthetic data set and from 1520 to 1368 for the natural data set from Peterson and Barney.

Vowel classification was by means of the linear discriminant analysis procedure described

in section 2.3.9. The results of the classification analyses for the two normalizing approaches for the PLURALITY synthetic data set previously described and the P & B data set are shown in Table 2.15. Note from the results that classification accuracy of the CLIH method

Table 2.15: Vowel classification using Neary (1977) classification scheme.

Data Set	Ref. Frame	N total	CLIH			CLIH3		
			# corr.	% corr.	APP score	# corr.	% corr.	APP score
PLURALITY	*EX*	1592	1379	86.6	.7513	1389	87.2	.7559
PLURALITY	*AV*	1592	1380	86.7	.7513	1389	87.2	.7559
P & B	P&B	1368	1199	87.6	.8063	1230	89.9	.8483

is only slightly improved with the CLIH3 method. In addition, note that the results utilizing the *EX* and *AV* reference frameworks are virtually identical. The lack of differences between these two approaches may be explained by the fact that while the tokens used for normalizing parameters in each approach appear quite different in location, the averages of the formants for each set represent points which are separated by only 0.02 log units in the *APS*. Thus it appears that either set may be used to yield reasonably high classification accuracy. However, if the results of this classification scheme are compared to the results of linear discriminant analyses on the synthetic data using $F1, F2, F3$ and x', y', z' previously shown in Table 2.7, we find no improvement in classification accuracy. This lack of improvement suggests that the extrinsic reference frameworks do not adequately reflect the data, or that, as was speculated initially in this section, this type of classification scheme is not appropriate for certain synthetic speech experiments.

2.5 Summarization and Discussion of Experiment I

In summary, this experiment has demonstrated that perceptual target zones for the vowels of American English can be constructed which span an extensive range of possible vowel

sounds. Additionally, these target zones are abutting and non-overlapping and correctly classify over 99% of the 1725 synthetic vowel-like sounds used in the experiment, based on identifications representing the plurality of subject's responses.

Comparisons of subjects' identifications indicated that subjects agreed with one another on about an average of 63% of the tokens and totally agreed with one another on about 19%. It is difficult to evaluate these percentages of agreement, since no attempt was made to determine how many tokens representative of non-American English vowels, ambiguous vowel sounds, or non-vowel sounds were included which could greatly influence identification agreements. As a reference, however, results from past studies utilizing identifications of natural vowels spoken in isolation reflect higher, but highly variable subject agreements, with identification error rates ranging from 43% (Strange, Verbrugge, Shankweiler, and Edman, 1976) to 3% (Kahn, 1978). While this large range may reflect differences in the stimulus parameters and experimental methodology utilized, these studies generally utilized stimuli which were produced with the clear intention of representing only salient American English vowels. Such a statement cannot be made for the present experiment.

Subjects agreed with themselves on identifications of about 75% of the tokens, a significantly greater amount than that found for between-subjects agreements. In addition, the identifications for some tokens of individual subjects consistently and confidently deviated from the plurality identifications suggesting that perceptual boundaries between vowels may reliably vary among individual listeners. This finding poses a problem for the development of zones to generically represent vowel classification by all listeners and may require the addition of other speech parameters (i.e., some "top-down" processing) to disambiguate token classification at zone boundaries.

If the number of responses comprising a plurality identification, the plurality frequency, is considered a measure of vowel saliency, we find that target zones are graded along this dimension, with the highest frequencies (i.e, most salient tokens) located generally more central to the zones and progressively lower frequencies (i.e., less salient tokens), toward zone boundaries. While this result may not come as unexpected, given the results of numerous studies exploring vowel boundaries with single continua, it does provide a more complete picture of the saliency gradients associated with the total areas of vowel cate-

gories and suggests that such gradients could be of some benefit applied to current methods of speech identification. Furthermore, while confidence ratings for identifications may have the potential to provide additional information to saliency gradients, the results of this experiment indicate that the subjectiveness of such a scale must somehow be reduced before such information can be obtained.

Comparison of the new target zones based on synthetic speech with current estimates of similar zones based on natural speech suggest that, while both sets of zones exhibit considerable amounts of overlap between like categories, both also demonstrate rather poor classification accuracy of data other than that utilized in their construction. The inaccuracy found for the synthetic-speech-based zones in classifying natural speech data may be attributable to the coarse resolution utilized in the mapping procedure. This resolution precludes the estimation of precise boundaries between vowel categories which would be required for higher classification accuracy and, as was preliminarily demonstrated, zones mapped at higher resolution can be considerably more accurate. Additionally, duplicating the subtleties of natural speech with synthetically-generated speech to a degree sufficient to assure that all elements of natural speech relevant to perception are intact is still a problem which cannot be dismissed.

Several reasons may account for the inability of the natural-speech-based target zones to provide high classification accuracy of the synthetic data. The first of these reasons reflects the original arguments that motivated the present experiment, that is, the utilization of insufficient amounts of data for uniform mapping of the vowel space combined with the uncertainty of using data from sources varying in analysis, formant measurement, and identification methodology, make zones estimates calculated in this way subject to noise and constant modification with the addition of new data. To maximize classification of the natural data used in their construction and minimize overlap between adjacent zones, the boundaries for these zones in the $x'y'$ dimension have necessarily become quite intricate and complex. Much of this complexity is due to the inclusion of outlying data points, even though much of the enclosed zone space is unaccounted for with data. A viable question thus becomes whether or not these zones should be modified to reflect at least some of the synthetic results.

An additional problem in classification of the synthetic data with the natural-speech-based target zones is their current inability to capture changes in zone boundaries related to the z' axis. Although this is a solvable problem, the zones are not currently able to capture shifts along the z' axis of constant values of $F1$ and $F2$ which tend to result in like vowel identifications, nor the more subtle shifts in zone boundaries which can result with changes in $F3$. A more detailed estimation of these zones in the z' dimension may not only descrease the complexity of boundaries as viewed in $x'y'$ planes, but also improve classification of the synthetic data.

The last reason to be discussed relevant to classification inaccuracies is a shared problem to both sets of target zones. This reason is based in the difficulties arising from the mismatch in the number of zones or vowel categories to be represented. While information concerning the locations and perceptual saliency of zones for [OW] and [EY] are of interest, these categories are not represented with natural-speech-based zones and have rarely been used for perceptual studies of vowel classification for American English in the past. Had the [OW] and [EY] categories been eliminated from the response set in the present experiment or included as zones for natural speech, we could anticipate considerably less confusion between these categories and their neighbors resulting in higher classification agreement.

Statistical classification procedures suggest that the plurality identifications of non-retroflex vowels in this experiment can be well accounted for by the frequencies of $F1$ and $F2$ and that the addition of $F3$ does not increase the relative classification accuracy unless the retroflex vowel category [ER] is included. While this finding suggests that $F3$ may perceptually function as a 'retroflexion detector' only, additional analyses of trends in subjects' identification agreements relative to $F3$ suggest that $F3$ does influence the saliency of vowels. In general, tokens with values of $F3$ in the range found for natural vowels are agreed upon by subjects to a greater extent than are tokens with values of $F3$ outside this range. Despite this probable influence on vowel saliency, $F3$ does not appear to greatly affect the identifications of non-retroflex vowels in American English. This finding casts some shadow of doubt on the necessity of representing $F3$, and in the case of the *APS*, a third dimension, in vowel classification. However, there remain several reasons for retaining $F3$ representation, despite its questionable utility for American English.

The first of these reasons is that, although boundary shifts in vowel categories due to changes in $F3$ are small, they do exist and thus may be best captured in multi-dimensional representations of the vowel categories. Small shifts in the perceptual boundaries between vowel categories related to changes in $F3$ have been reported by Fujisaki and Kawashima (1968), Holmes (1986), and Neary (1989). Additionally, it can be demonstrated that boundary shifts related to changes in $F3$ also occurs in the present experiment. If the locations of all tokens identified as non-retroflex vowels from the present experiment are recalculated to reflect a unitary value for $F3$, a single set of two-dimensional zones for one z' plane can be utilized for their classification. Errors found in this classification should reflect in a general sense any boundary shifts induced by changes in $F3$, since the multi-dimensional zones constructed in the present experiment can correctly classify 100% of these identifications when $F3$ varies. The determination of which z' plane of zones may best represent all tokens should not be extremely critical if the assumption that boundaries in terms of $F1$ and $F2$ values do not shift with changes in $F3$, although intuitively the z' plane most representative of $F3$ values found in natural speech seems implied. Since the average z' value for non-retroflex vowels in the natural speech data considered in Table 2.8 is 0.70, the zones associated from that plane will be utilized for the classification. The value of $F3$ for all tokens (excluding rejected tokens) identified by the plurality of subjects as non-retroflex vowels was set to 2528 Hz, forcing their z' location to 0.70, and values for their locations in x' and y' recalculated. Classification of these tokens with the synthetic-speech-based zones for the $z' = 0.70$ plane was 88.3%. This result suggests that identifications of almost 12% of tokens may have been perceptually influenced by the value of $F3$ and that, while multi-dimensional zones can easily take such influences into account, two-dimensional zones based on values of $F1$ and $F2$ may have difficulty accomodating perceptual shifts influenced by $F3$.

An additional reason for $F3$ representation takes into consideration the more global potential for a third dimension. A metric reflecting $F3$ information is able to provide the potential for establishing target zones, and therefore classification, of languages other than American English. In particular, classification of languages which necessitiate perceptual differentiation of the rounded/unrounded distinction for vowels may require $F3$ representa-

tion. The need for representing $F3$ information to increase vowel classification over $F1$ and $F2$ information alone has been described by Fant (1973) for Swedish and Pols, Tromp, and Plomp (1973) for Dutch. The necessity of $F3$ information, as reflected in the z' dimension of the *APS*, for the disambiguation of rounded and unrounded vowels in German has been demonstrated in Jongman, Fourakis, and Sereno (1989).

An additional consideration for representing $F3$ information in vowel classification is its potential as a normalizing factor for differences between talkers. Generally, the variance found in $F3$ for vowels (rhotacized vowels excepted) of an individual talker is quite small compared to the variance found across talkers (Fujisaki and Kawashima, 1968). Furthermore, $F0$ and $F3$ are usually highly correlated for an individual talker, reflecting the generally close relation between the sizes of the glottis and vocal tract. These facts have been employed in a number of studies investigating how changes in $F0$ and $F3$ may affect vowel perception and how information from these two parameters may be used in across-talker normalization strategies for vowel classification (See Neary (1989) for a discussion). While results from these studies are often conflicting, a consistent finding is that perceptual boundary shifts between vowel categories are greater with concomitant changes in $F0$ and $F3$ than with either alone (Fujisaki and Kawashima, 1968; Neary, 1989). However, the issue of how to best utilize both $F0$ and $F3$ in a talker normalization strategy remains largely unresolved. The range of $F3$ utilized in the present experiment is considerably greater than is likely to be found associated with an individual male $F0$ in natural speech, thus perhaps creating an environment too unnatural for evaluating target zones intended for natural speech. The talker-normalization parameter utilized in the *APS*, *SR*, is currently influenced predominatly by $F0$ in its calculation, although through its defined intention to represent a talker's vocal characteristics, could represent additional parameters, like $F3$, should normalization by these means prove superior.

Another relevant issue concerning the role of $F3$ in vowel perception is rhotacization, that is, the auditory property of "r-coloring" of vowels. While the vowel [ER] may be considered completely rhotacized, as in the word "bird", partially rhotacized, or r-colored, vowels also occur in American English, as in the words "beard, bared, bored." As was mentioned previously in Section 2.3.1, subjects had difficulty labelling tokens which were

neither clearly non-retroflex monophthongs nor [ER], implying that these tokens may have been representative of r-colored vowels. However, the partial rhotacization of vowels is generally considered a dynamic process, whereby the intitial vowel sound is non-rhotacized and becomes increasingly more rhotacized over time (Ladefoged, 1982, p. 207). Since there was no formant movement in the present stimuli, the tokens in question do not fit the general description of r-colored vowels, but rather, may represent steady-state versions of formant patterns associated with single points in time along the dynamic of the partial-rhotacization process. While it is not likely that such sounds would be found in natural speech, the acoustic and perceptual aspects of r-colored vowels in American English has not been extensively investigated. The *APS* can provide an excellent framework in which to base further investigation.

Several schemes for natural speech vowel classification were compared and contrasted with the synthetic-speech-based target zone approach in terms of classifying the plurality-based identifications from this experiment. Each of these other schemes had difficulties with this task due to either inherent flaws in the classification approach itself or theoretical constraints to natural speech which made them potentially inappropriate for classification of some synthetic speech. The $F1$ x $F2$ ellipses from Peterson and Barney (1952) are flawed in that identifications related to changes in $F3$ cannot be reflected. Previous discussion has related the importance of $F3$ representation in vowel classification. The classification scheme from Syrdal and Gopal (1986) based on patterns of binary vowel features related to bark-transformed formant differences is flawed in that additional speech parameters like vowel duration are required to disambiguate vowel categories. While such parameters may prove capable of aiding the classification of natural speech with this approach, these parameters are not available for the synthetic tokens utilized here. The last classification scheme evaluated, from Neary (1977), requires the extrinsic specification of information about the distribution of vowels for an individual talker to provide a reference framework for classification. Although across several approaches to establishing such a framework with the synthetic identifications were attempted, classification results utilizing these frameworks were no better than for statistical classification with simple formant values. These results suggest that such a classification scheme may not be appropriate for use with synthetic

speech experiments where tokens potentially do not represent those of an individual talker.

Chapter 3

Experiment II: Estimation of difference limen for distance (d) in the APS vowel space

3.1 Introduction

With the general areas of *PTZs* for synthetic vowels in the *APS* established in Experiment I, a more detailed investigation of the locations of the boundaries between PTZs is required. However, before such a task can be undertaken, an additional question of some importance must be addressed. While we can surmise from Experiment I that .05 log unit resolution is too coarse for accurate boundary estimation since boundaries estimated with .01 log unit resolution from a phonetically trained subject increase correct classification of natural vowels substantially (Hawks and Miller, 1989), mapping the entire vowel space utilized in Experiment I at this resolution would increase the number of tokens to be judged by an order of magnitude. For example, at the .05 log unit resolution employed in Experiment I, 304 tokens were evluated in the $z' = .70$ plane. Mapping this same plane at .01 log unit resolution requires 7703 tokens. In addition, such a mapping may potentially reflect considerable amounts of redundant information should neighboring tokens be perceptually identical. Thus the question of what level of resolution is required for sufficiently defining

PTZs and their boundaries in the *APS* requires attention. One method of answering this question is to determine the difference limen (*DL*) for distance (*d*) in the *APS*.

The distance (*d*) between any two points specified in the *APS* may be given by the equation,

$$d = ((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2)^{1/2} \tag{3.1}$$

where x_1, y_1, z_1 and x_2, y_2, z_2 are the coordinate values of the two points in *APS*. Data on difference limens (DLs) for formant frequencies of vowel sounds (Flanagan, 1955; Kakusho and Kato, 1968; Mermelstein, 1978; Nord and Sventelius, 1979) indicate that an average DL for *d* should be on the order of 0.02 log units. However, when extreme DL values for the first and second formant frequencies of vowel-like sounds from studies by Flanagan (1955) and Mermelstein (1978) are converted to *APS* coordinate values, they yield values for *d* which vary from as large as 0.0947 to as small as 0.0056 log units. Additionally, it should be noted that neither of the aforementioned studies varied $F3$ and only Mermelstein (1978) investigated simultaneous variation of formant frequencies as a sub-condition of one experiment. Thus, these findings may not reflect the difference limen found when more than one formant frequency is varied at a time. In his concluding remarks, Flanagan (1955) called for the experimental work to be detailed here, i.e., a mapping of DL areas on the $F1 - F2$ plane and the simultaneous variation of multiple formants, as a necessary extension of his work reported at that time, although, to the author's knowledge, no such work has been reported. Although parametrically varying the frequencies of up to three formants simultaneously is difficult to specify in a common metric, the *APS* provides an excellent format for such specifications since the complicated changes reduce to interpretable changes in location. Given consideration of these factors, we are motivated to determine the *DL* for *d* through our own investigation.

An additional relevant issue in speech perception research is whether or not discrimination ability is greater for vowel stimuli belonging to different phonetic categories as opposed to stimuli belonging to the same category. Fry, et al. (1962) found no significant differences between the discrimination of isolated vowels, whether from within a phoneme region or spanning a phoneme boundary. Stevens, et al.(1969), on the other hand, found distinct peaks in discrimination functions at vowel boundaries. Pisoni (1973) also found evidence

for differences in vowel discrimination of within-category versus between-category stimuli related to the comparison delay interval. Macmillan, Braida, Goldberg, and Khazatsky (1986) suggest, however, that many of the discrimination differences reported can be explained in terms of the stimulus range and psychophysical method employed and that these differences may not exist when proper stimulus parameters and methods are employed.

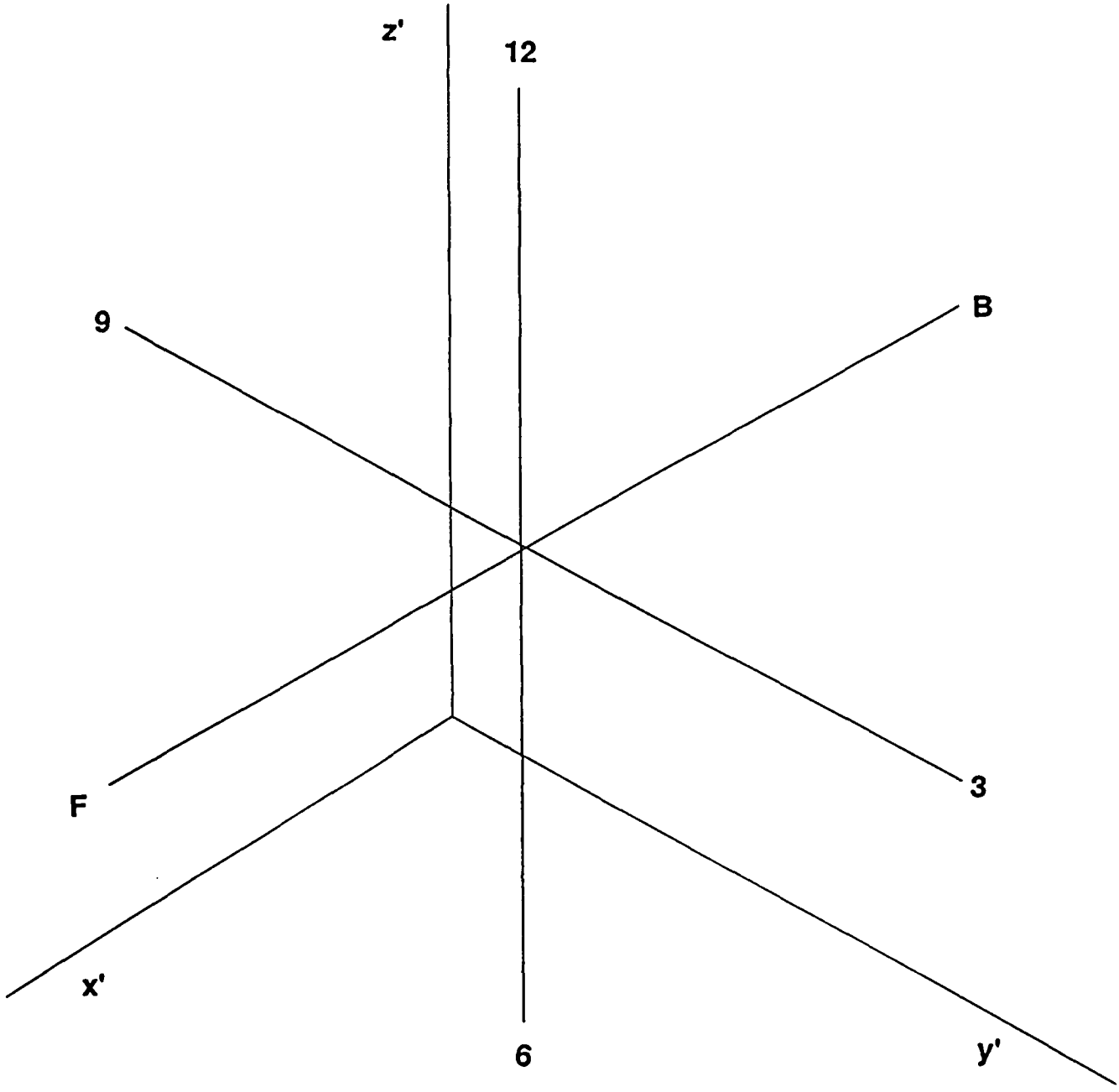
In Experiment I, a roving identification paradigm was used in an effort to restrict subjects to use only their own internal references for vowel sounds as a judgement basis for their responses. While this method does not yield as sensitive results as fixed discrimination paradigms (Macmillan, et al., 1986), it is considered adequate for the large range and relatively coarse resolution required in that experiment. However, a higher level of discrimination may be required for determining more exact boundary locations. Macmillan, et al. (1986), in generalizing the Durlach and Braida (1969) theory of intensity perception to speech perception, indicate that resolution performance is highest when 1) the stimulus range is small, 2) the inter-stimulus interval is small, and 3) a fixed discrimination paradigm is used, such as two-interval forced-choice (2IFC) or same-different (AX). The design of Experiment II reflects consideration of these factors.

3.2 Methods

3.2.1 Stimuli

Vowel tokens were synthesized using formant frequencies specified by points along continua in the *APS*. These continua represent straight lines in the *APS*, with groupings of three continua sharing a common center value, or *reference*. Each three-continua grouping is oriented so as to graphically form a three-dimensional "cross-hair", as shown in Figure 3-1, with each continuum of the crosshair parallel with one of the x', y', z' axes. Recall that these axes are transformations of the *APS* dimensions, x, y, and z (See Section 1.2.1). Each continuum is further subdivided at the reference point, such that the reference serves as a common point shared by each of what is now six continua. For convenience, the six continua are labeled in clock-like fashion for continua parallel to x' and y' (i.e., 12 o'clock= upward along y', 6 o'clock= downward along y', 3 o'clock= right along x', and 9 o'clock= left along

Figure 3-1: Orientation of the six continua in APS $x'y'z'$ coordinates associated with each reference point used in Experiment II.



x'), and F and B (i.e., front and back) for continua parallel to z' (See Figure 3-1).

Initially, a fixed length of 0.02 log units and step size of 0.00025 log units was evaluated for each continuum which yielded 80 tokens per continuum. However, as will be discussed later, the amount of formant change given a fixed step size varies with the axial orientation of the continuum. Thus, given that the synthesizer requires integer formant value specifications, the initial step size utilized sometimes yielded identical formant values for neighboring tokens. Additionally, early pilot studies revealed that the initial continuum length may be too short to provide adequate estimation of some difference limens. Based on these studies, the following specifications were utilized to alleviate these difficulties and aid in the efficiency of testing. Continua parallel to the x' axis were 0.02 log units long with 80 tokens spaced at 0.00025 log units, continua parallel to the y' axis were 0.04 log units long with 120 tokens spaced at 0.00033 log units, and continua parallel to the z' axis were 0.015 log units long with 60 tokens spaced at 0.00025 log units. Additionally, a computer program was implemented which evaluated the formant specifications for all tokens in a given continuum and eliminated any tokens which were duplicates. Although the continua were now of varying lengths and step sizes, these changes assured that the difference limens could be estimated within a reasonable range and that all tokens along any given continuum varied by at least 1 Hz in at least one of the first three formant specifications.

Token duration, *F0* and amplitude contours, formant-bandwidth calculation, higher-formant values, and all other global synthesis parameters were identical to those previously specified for token synthesis in Experiment I (See Section 2.2.1).

Reference points were selected at 17 locations in the *APS*, 10 from the interiors of the target zones, [IY, IH, EH, AE, AA, AH, AO, UH, UW, ER] and 7 from estimated boundary areas between the target zones, [IY-IH, IH-EH, EH-AE, AE-AH, AH-AA, AH-UH, UH-UW]. This design yielded a total of 102 continua for evaluation.

Reference point selection criteria

Since this experiment began prior to the analysis of the results from Experiment I, locations for the 10 reference points from the interiors of target zones, hereafter referred to as *center* references, were selected based on the results of an earlier pilot study. This pilot study was

identical to Experiment I and utilized three subjects highly trained in phonetics. Tokens whose identifications had been twice unanimously agreed upon by the three subjects were first located for each of the 12 response categories used in Experiment I. If more than one token per category satisfied this criterion, the final token selection was made by the investigator. All center reference points were located in the $z' = 0.70$ plane except for [IY] ($z' = 0.75$) and [ER] ($z' = 0.55$). The locations of the center references ([ER] is not included) are graphically illustrated in x', y' space in Figure 3-2, along with the synthetic target zones for the $z' = 0.70$ plane from Experiment I. Additionally, the locations of the "EXEMPLARY" tokens (See Section 2.4.2) from Experiment I and the male averages from Peterson and Barney (1952) are shown for comparison. All points in the figure not originally located in the $z' = 0.70$ plane have been normalized in a manner previously described for Figure 2-18 in Section 2.4.4 such that their relative locations reflect only their values of $F1$ and $F2$.

Locations for the 7 reference points from boundary areas between zones, hereafter referred to as *ambiguous* references, were selected in a manner different from the center references. Gross areas of potentially ambiguous tokens in the $z' = 0.70$ plane for the seven boundaries were first selected based on identifications of synthetic vowels from a previous study (Miller and Hawks, 1989) described in Section 2.4.3. The tokens in these areas had received identifications that were mixed between the vowel categories they bordered, as well as generally low confidence ratings. Points for the tokens in these areas are shown in Figure 3-3 along with the target zones constructed for this z' plane from Experiment I. Further refinement of these areas was accomplished by examining the identifications of these same tokens from three new subjects, highly trained in phonetics. Once again, tokens receiving identifications mixed between the appropriate neighboring categories as well as low confidence ratings were noted. These tokens are shown in Figure 3-4. The final selection of reference points was made from these groups of tokens by the investigator and are shown in Figure 3-5.

Figure 3-2: Locations in APS $x'y'$ coordinates of center references (x's) utilized in Experiment II compared to locations of exemplary reference framework tokens (+'s, See Section 2.4.2) and male average vowels ($\star$'s) from Peterson and Barney, 1952.

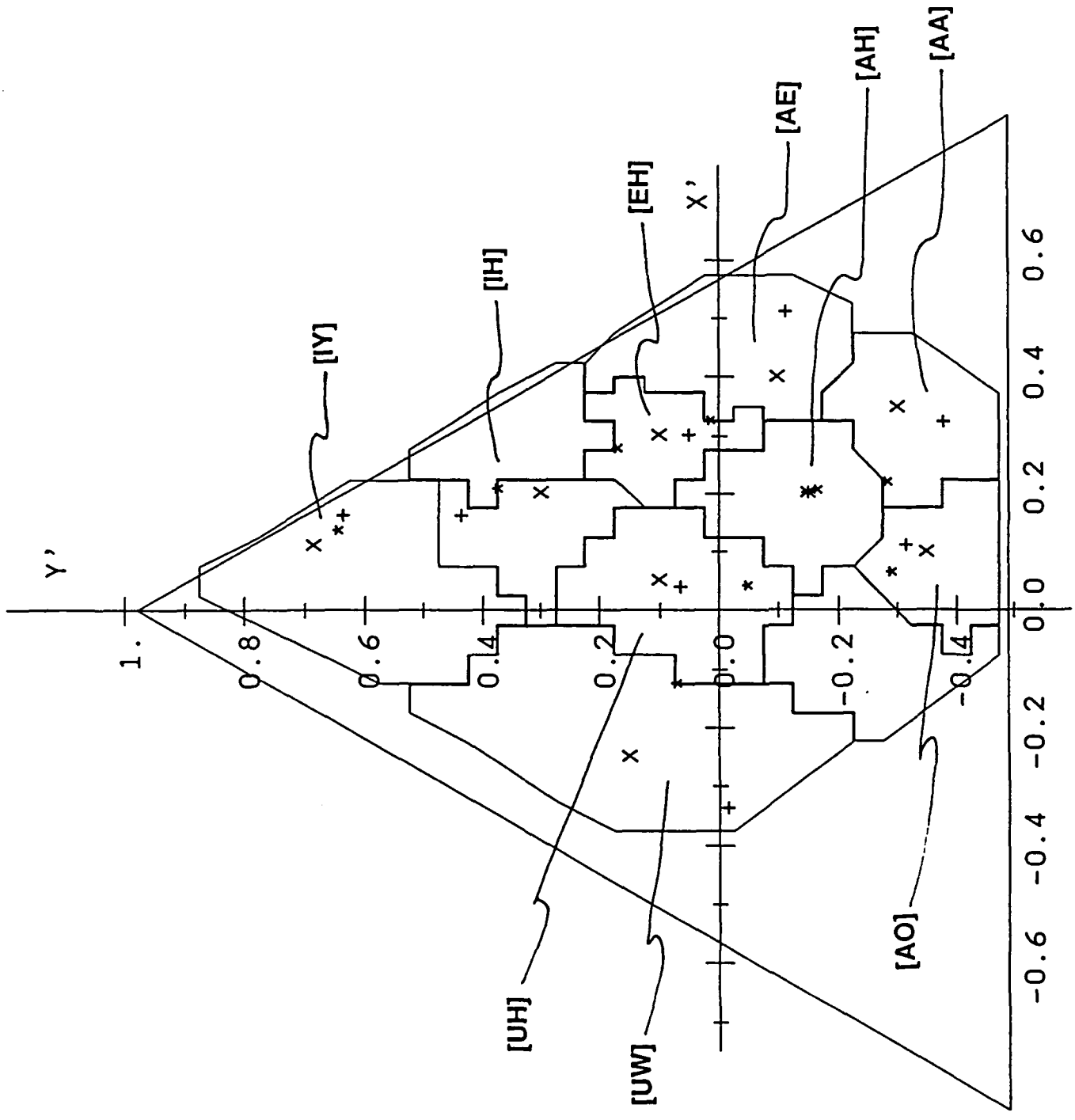


Figure 3-3: Locations in APS $x'y'$ coordinates of points for the first evaluation for ambiguous reference points along with *SSB* target zones for the $z' = 0.70$ plane.

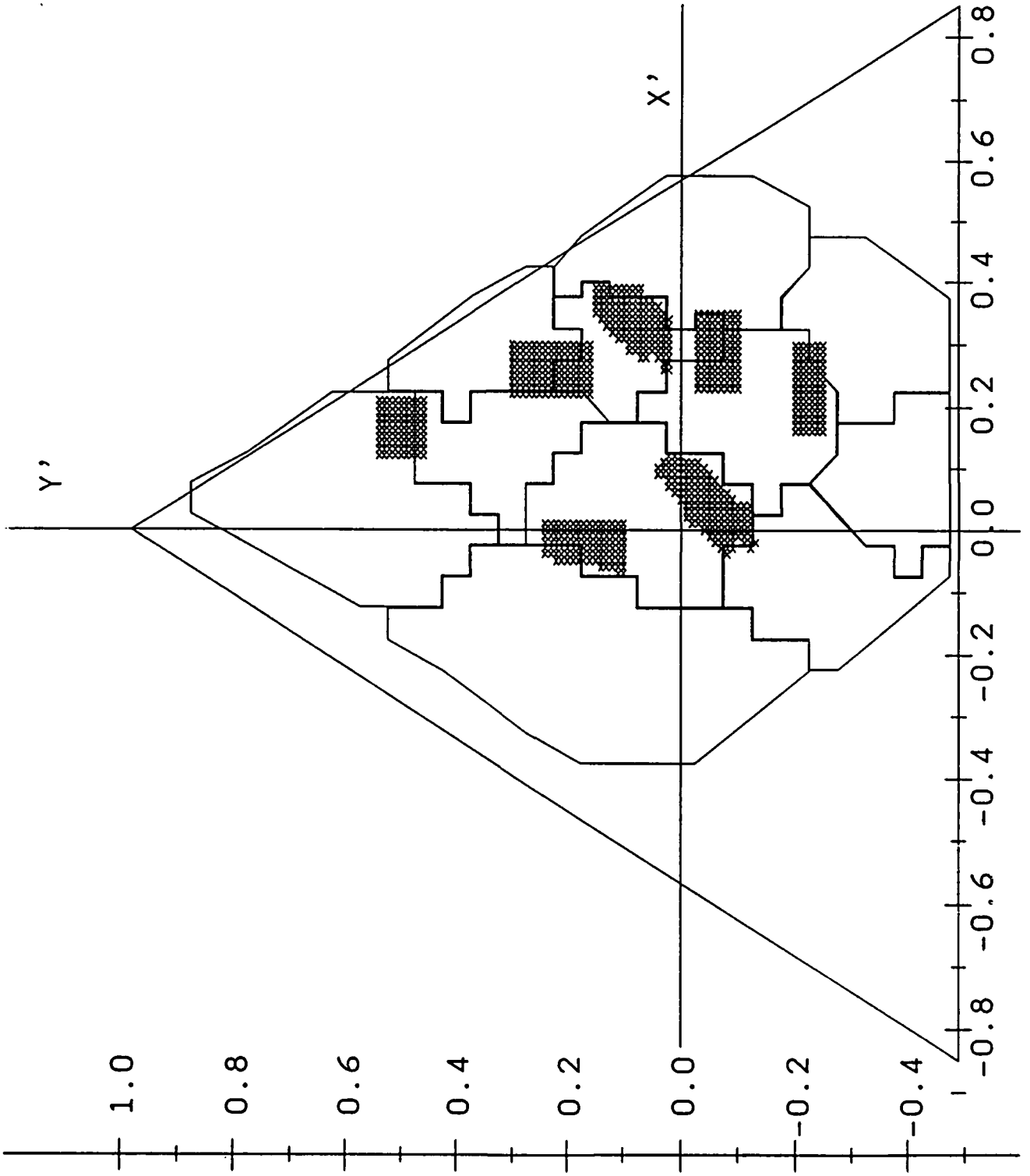


Figure 3-4: Locations in APS $x'y'$ coordinates of points for the second evaluation for ambiguous reference points along with *SSB* target zones for the $z' = 0.70$ plane.

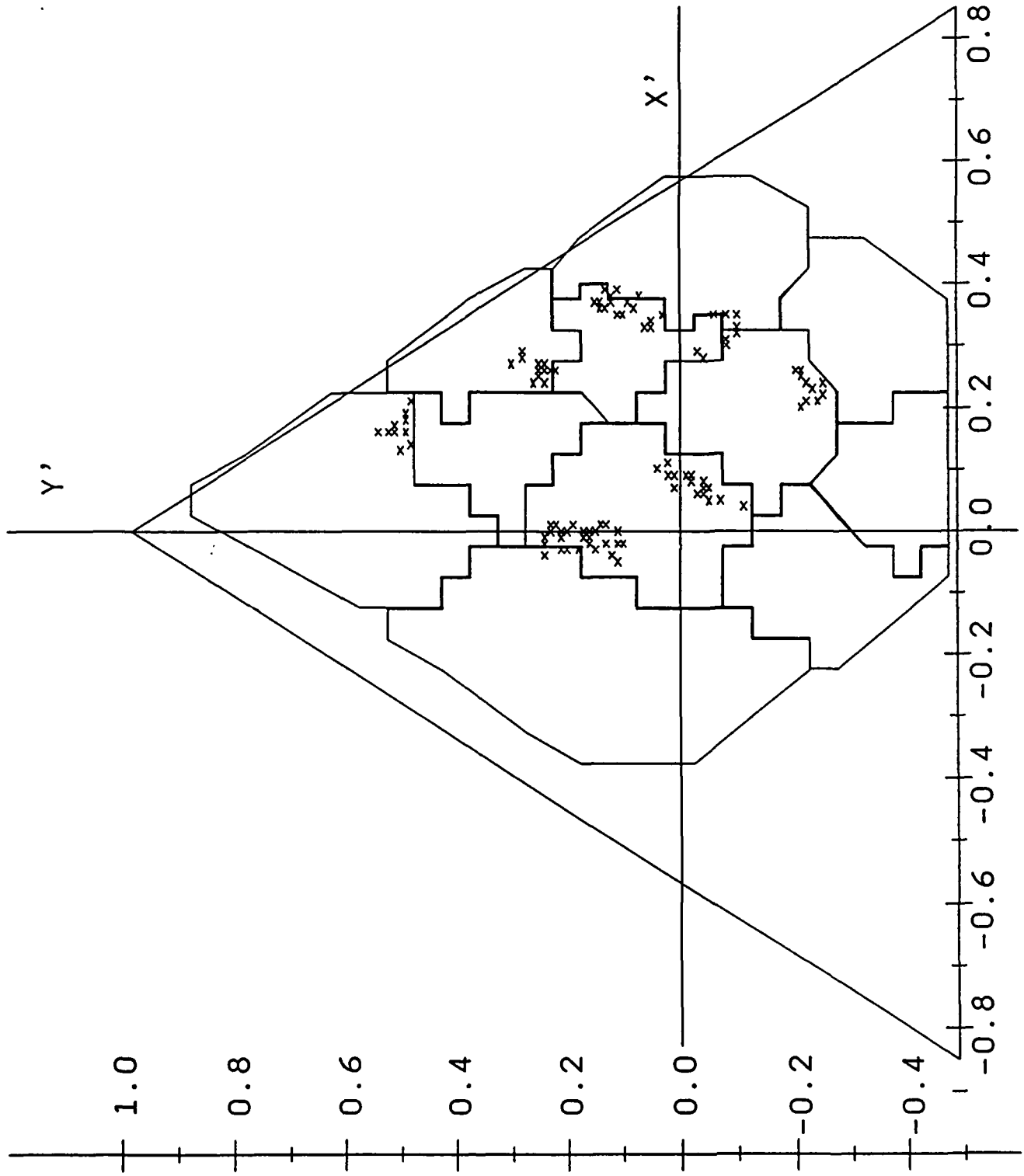
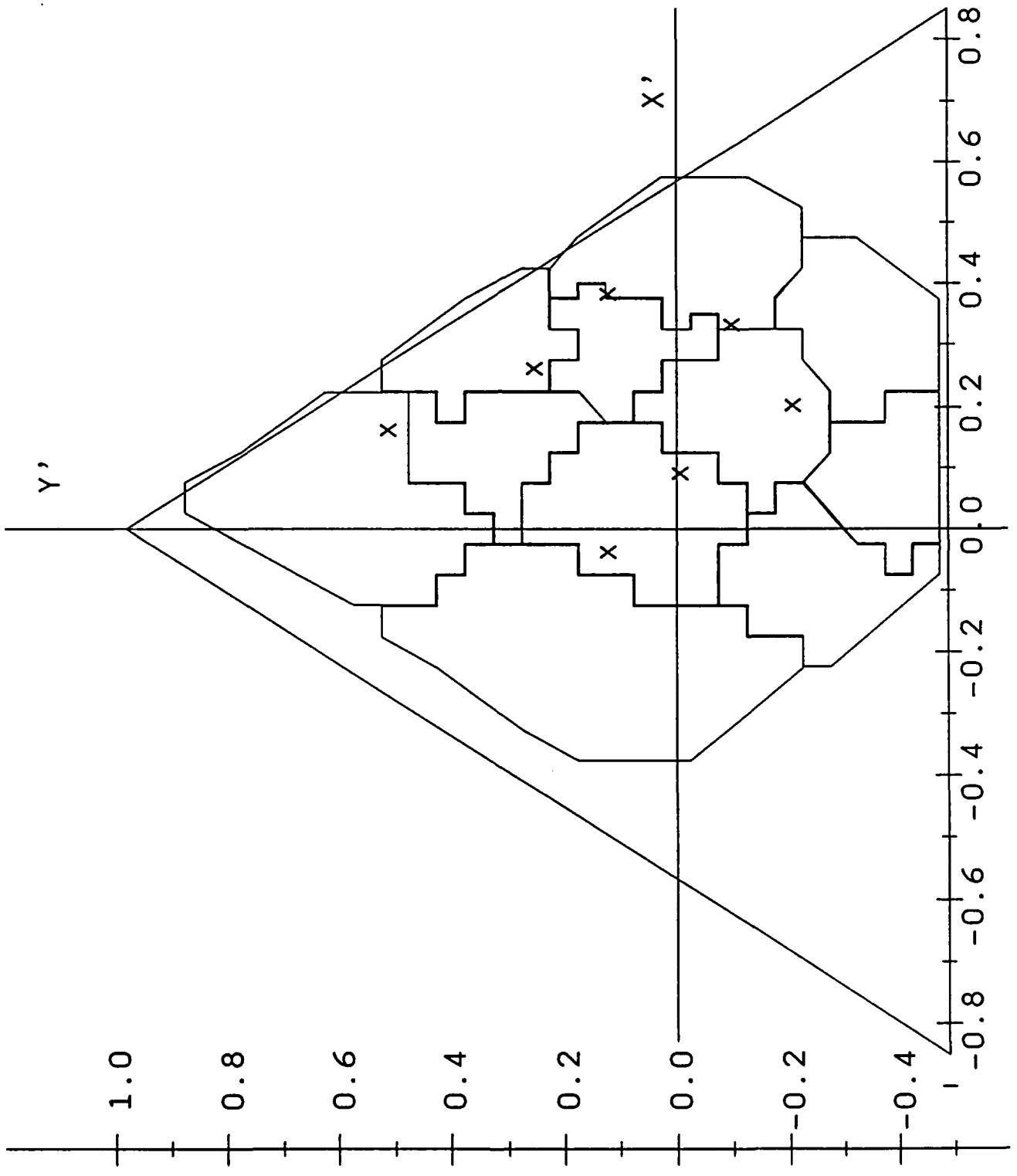


Figure 3-5: Locations in APS $x'y'$ coordinates of the ambiguous reference points used in Experiment II along with the *SSB* target zones for the $z' = 0.70$ plane.



3.2.2 Procedure

An adaptive up-down procedure similar to the PEST procedure (Taylor and Creelman, 1967) was utilized employing a cued, two-alternative forced-choice (2AFC) task. In this procedure, stimuli were presented in decreasingly smaller steps with the directional movement along the continuum determined by the subject's previous responses. The initial step size was 20 and approximately halved the first four times the direction along the continuum reversed (10, 5, 2, 1). Unlike the regular PEST procedure, no provision was made for increasing the step size. A 3-down, 1-up rule was also employed to the directional criteria. This rule allows movement toward the reference only after 3 correct responses at the current level and movement away from the reference after 1 incorrect response. This strategy should yield a probability of a correct response at the point of convergence of 0.794 (Levitt, 1970).

A fixed-standard vowel token was presented as a cue followed by two other vowel tokens, one of which was the same as the standard. The three vowel tokens comprising each trial were separated by inter-stimulus intervals of 250 msec. Subjects were asked to judge whether the different sound was in the first or second interval following the cue. The fixed-standard vowel tokens used represented the reference points of the cross-hairs. Thus, the six thresholds from each cross-hair were estimated by measurements made from the same reference point outward in all six directions. The response interval was subject paced, such that the next trial began 3 sec after a response was registered. Trial presentations ended, constituting a block of trials, after the subject had made 14 reversals along the continuum. Data from the first four reversals were discarded and the threshold was estimated as the average of the remaining 10 reversal points.

Subjects worked two hours per day and could complete about twelve blocks of threshold measurements within this time period. Blocks were randomized among the 17 reference points and the six directions for each subject. Each continuum was evaluated twice by each subject, yielding a total of 816 (17x6x4x2) threshold estimates.

Although same/different (AX) tasks have often been used for experiments of this type (Flanagan, 1955; Mermelstein, 1978; Nord and Sventelius, 1979), the cued, two-alternative forced-choice task is selected here in an effort to eliminate response criterion differences

between subjects. With a fixed-standard cue and the requirement of a judgement based on the comparison of two intervals, subjects are less likely to utilize an internal standard. A 250 ms delay interval has been found to yield the highest values of d' in discrimination experiments with vowels (Pisoni, 1973), and was therefore used here as the inter-stimulus interval.

3.2.3 Apparatus

The vowel tokens were synthesized at a 10 kHz sampling rate and stored on a DEC 3200 computer. The testing paradigm was implemented as an interactive program on the same computer which each subject ran independently. The program operated as follows. Prior to each trial, a warning flashed on the screen that a new trial would begin in 2 seconds. Following the trial presentation, subjects saw a question on the screen asking whether the different vowel sound was in the second or third interval. Subjects entered their response by pressing one of two labeled keys on the keyboard. A file was generated for each block of trials which contained general subject and block information, as well as a detailed description of the presentation parameters and subject's responses for that block of trials.

The stimuli were presented via a MicroTechnology Digisound-16 digital-to-analogue converter followed by a passive 5 kHz anti-aliasing filter. Subjects heard the stimuli binaurally over Sennheiser HD-430 headphones at a comfortable listening level (55-60 dBA-Slow SPL) in a sound-isolated room.

3.2.4 Subjects

Subjects, two male and two female, were recruited from the student body of Washington University and the nearby St. Louis area. Subjects' ages ranged from 17 to 25 years. All subjects were native speakers of American English with no known history of either speech or hearing impairment. All subjects were naive in terms of any formal phonetic training.

3.2.5 Training

Training consisted of a minimum of four runs each on 14 continua not used in the actual experiment. Subjects were rejected if, by the end of the four runs, the average DL for

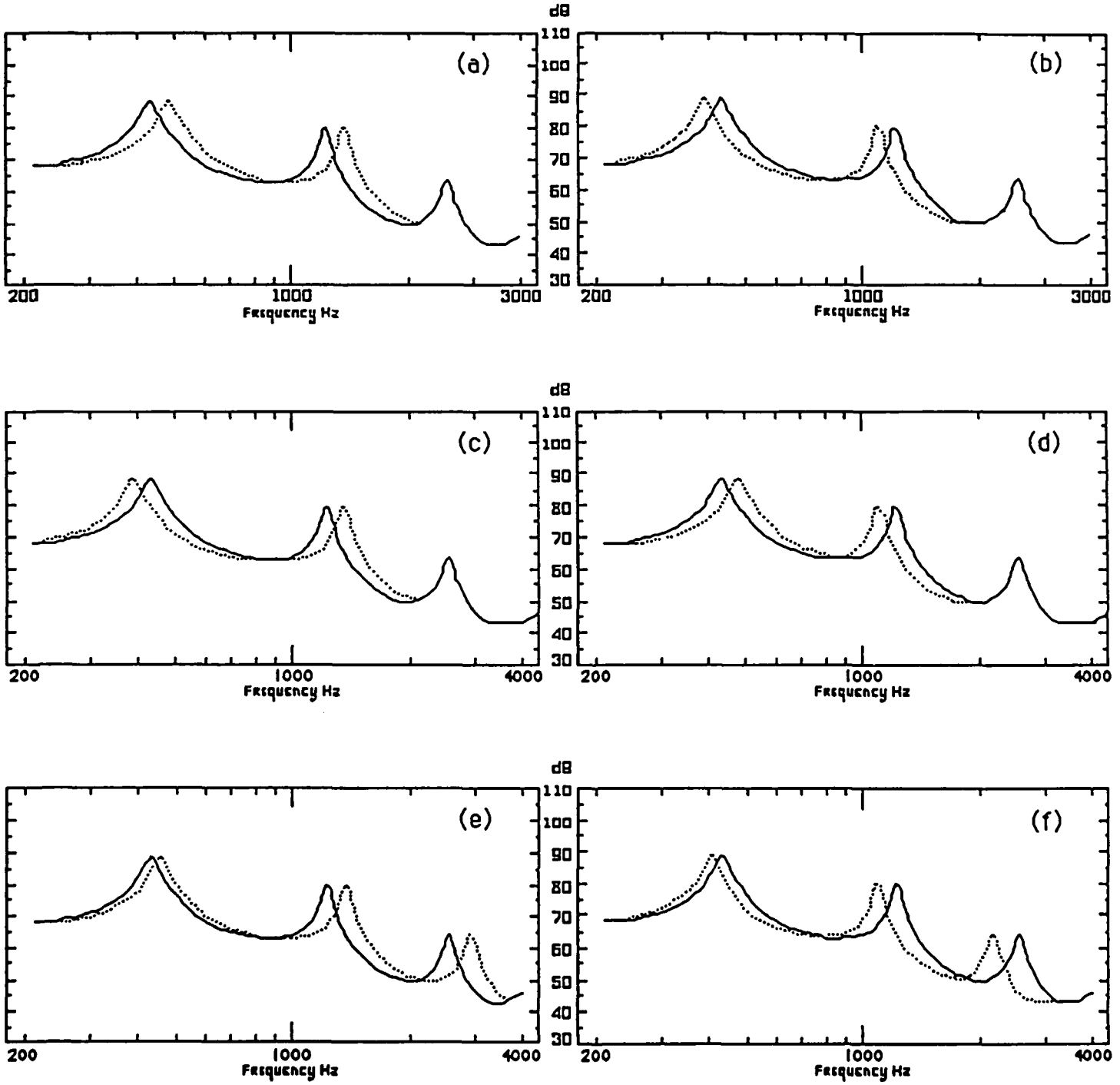
d over any one set of the 14 continua was not equal to or less than 0.01 log units. One subject was eliminated and replaced due to this criterion. The continua utilized in training included the six directions associated with each of the [OW] and [EY] center references and two additional continua, one from a central point in [AH] and one from the [AH-AO] boundary area. All other procedures and methods were identical to those used in the actual experiment.

3.2.6 Formant change and *APS* continua

To better describe the changes in spectral pattern under evaluation, this section will discuss the changes in formant frequencies derived from movement along the *APS* continua utilized in this experiment. First, consider points in *APS* at a fixed distance of 0.02 log units away from a cross-hair reference point along each of the six continua for that cross-hair. Given a fixed fundamental frequency, these six points and the reference point can be thought of as representing vowel sounds which differ systematically in two or all three of their specifications for $F1$, $F2$, and $F3$. These differences in formant specifications may be expressed as the differences between the logarithms of the formant frequencies represented by each of the six points along the continua and the logarithms of the formant specifications for the reference point. We will find that formant frequency shifts for $F1$ and $F2$ along some continua are often equal when expressed in this manner. However, a more traditional measure, $\Delta F/F$, where ΔF is the difference between the continuum point and the reference point formant frequency and F is the reference point formant frequency. This measure, expressed as a percentage, is also approximately equal for the relatively small changes we will be considering here and will therefore be used extensively for analyses of the results.

Continua 3 and 9 (See Figure 3-1) form a straight line through the reference point and lie parallel to the x' axis. Moving the fixed distance away from the reference point along continuum 3 yields positive changes in both $F1$ and $F2$, which are equal log differences between continuum and reference point formant frequencies, and approximately equal 3.3% changes in $F1$ and $F2$. Thus, for this continuum, $F1$ and $F2$ increase proportionately together. These differences are idealized graphically in Figure 3-6a as spectra, with the reference point represented with a solid line and the point along continuum 3 with a dashed

Figure 3-6: Idealized spectra representing the relative shifts in formant frequency for a fixed distance movement along each of the six directional continua (dashed lines) relative to the reference point (solid lines). (a) Continuum 3; (b) Continuum 9; (c) Continuum 12; (d) Continuum 6; (e) Continuum F; (f) Continuum B.



line[1]. Moving this same distance away from the reference point along continuum 9 yields an opposite effect (See figure 3-6b) whereby $F1$ and $F2$ both decrease proportionately by 3.3%. There is no difference between the $F3$ specifications of the reference point and points along these continua.

Continua 6 and 12 form a straight line through the reference point and lie parallel to the y' axis. Movement along these continua does not yield parallel changes in $F1$ and $F2$ as seen in continua 3 and 9, but rather, opposing changes, whereby $F1$ and $F2$ shift toward or away from one another an equal log differences between continua and reference point formant frequencies. For movement equal to the fixed distance away from the reference point along continuum 12, $F1$ will decrease by approximately 1.9% of the reference point $F1$ value and $F2$ will increase by the same percentage (Figure 3-6c). Once again the opposite effect is found for equivalent movement along continuum 6 (Figure 3-6d), whereby $F1$ will increase and $F2$ will decrease by approximately 1.9% of the reference point formant values. There is again no difference between the $F3$ specifications of the reference point and points along these continua.

For movement along the continua parallel to z', F and B, a more complicated pattern emerges. Movement equal to the fixed distance away from the reference point along continuum F results in an increase in all three formant values relative to the reference point values. $F1$ increases by approximately 2.8%, $F2$ by 5.5%, and $F3$ by 8.3% (Figure 3-6e). The reverse of these changes occurs for the fixed-distance movement along continuum B with all three formant frequencies decreasing by approximately these same percentages (Figure 3-6f).

This section has related how vowel sounds represented by points along the continua of a cross-hair vary in their formant frequency specifications from the specifications represented by the reference point for that cross-hair. From this discussion several important aspects of formant change with movement in *APS* emerge. First, the differences in formant specifications associated with points along a straight-line continuum which is parallel to one of the x',y', or z' axes in *APS* are systematic and may be calculated given knowledge of the

[1]The shifts indicated for formant frequencies between spectra in these figures has been exaggerated for visual clarity.

distance between the points and the axial orientation of the continuum. Second, the axial orientation of the continuum also determines the nature of formant change which results in distinctly different patterns of spectral change associated with each axis. For further elaboration on the relationships between movement in *APS* and formant changes, see Appendix A.

3.3 Results

3.3.1 General Analyses

Table 3.1 shows the mean DLs (rounded to two significant digits) as log unit distances from the reference point by reference point group, axis, and axial direction across subjects and replications. The overall average DL expressed as d across all continua was 0.0110 log units. The average DL for continua parallel to the x' axis was 0.0091 log units, the y' axis, 0.0190 log units, and the z' axis, 0.0035 log units. A repeated-measures ANOVA was computed on the experimental results employing the four subjects as replicates. A 2 x 3 x 2 x 2 factorial design was used for the analysis comparing the two reference point groups (center vs. ambiguous), the three axes (x', y', z'), the two axial directions (positive vs. negative relative to the reference point), and the two replications. Subjects' average DLs within each reference point group for each factor served as the analysis data.

The statistical probabilities of a significant difference for each comparison factor from this analysis are shown in the second column of Table 3.2. The most significant main effect found in the analysis was for differences between DLs when grouped by axis ($\rho = 0.0002$). Given the previous discussion of the differences in percent change of formant frequencies for a fixed-distance movement along these axes, if the DLs expressed as percent formant change for the three axes are relatively equal, these differences in DLs expressed as distance are not surprising. The only other significant main effect found was for replication ($\rho = 0.02$). Subjects demonstrated significantly better performance on replications of blocks, despite training and randomization of blocks. No significant difference was found between center and ambiguious reference point groups, although the average DL for the ambiguous reference point continua is smaller than the average DL for the center reference point continua.

Table 3.1: Mean DL in log unit distance for various conditions across subjects and replications.

Axis	Axial Direction	*Center* References	*Ambiguous* References	Row $\overline{x}$
x'	+	.0092	.0080	.0086
x'	−	.0110	.0083	.0095
x'	±	.0100	.0082	.0091
y'	+	.0180	.0200	.0190
y'	−	.0220	.0180	.0200
y'	±	.0200	.0190	.0190
z'	+	.0040	.0033	.0036
z'	−	.0040	.0028	.0034
z'	±	.0040	.0031	.0035
$\overline{x}$	+	.0100	.0100	.0100
$\overline{x}$	−	.0120	.0096	.0110
$\overline{x}$	±	.0110	.0100	.0110

Additionally, no significant difference was found between directions along the continua (e.g., continua 12, 9, and F vs. continua 6, 3, and B).

Table 3.2: Probabilities of significance for factors from overall and individual reference group analyses-of-variance of DLs expressed as distance.

Factor	Overall	*center* references	*ambiguous* references
Reference Group	ns	$\rho = .0017$	ns
Axis	$\rho = .0002$	$\rho = .0001$	$\rho = .0010$
Direction	ns	$\rho = .0170$	ns
Replication	$\rho = .0200$	$\rho = .0430$	ns

Significant interactions were found for axis-by-replication ($\rho = 0.012$), reference group-by-direction ($\rho = 0.026$), and reference group-by-axis-by-direction ($\rho = 0.033$). The axis-by-replication interaction seems to reflect the fact that while subjects' DLs improved across replications for the x' and y' axes, their performance remained the same across replications for the z' axis. The reference group-by-direction interaction stems from opposing differences in DLs between directions for the two reference point groups. The mean DL for increasing axis directions is smaller than for decreasing axis directions for center reference point continua, while the opposite is true for the ambiguous reference point continua. The three-way interaction between reference group, axis, and direction is more complex, but appears to predominantly reflect the reference group-by-direction interaction and the extremely strong main effect of axis.

3.3.2 Analyses by reference group

As was stated previously, no significant difference between the center and ambiguous reference points was found when they were considered as two distinct groups. However, that analysis did not evaluate possible differences within the groups. Separate ANOVAs were computed on the ten center reference points and the seven ambiguous reference points utilizing the same grouping factors of reference, axis, direction, and replication, and employing subjects as replicates. The statistical probabilities of a significant difference for each com-

parison factor from these analyses are shown in the third and fourth columns of Table 3.2. A significant main effect ($\rho = 0.0017$) for differences between the ten center reference points was found, however, a similar effect was not found for the seven ambiguous reference points. Significant differences for the axis factor were found for both reference point groups, however, differences for the direction and replication factors were significant only for the center references group.

Center reference points

The overall DL results, ranked in order from smallest to largest, for continua from each of the ten center reference points are shown in Table 3.3. Additionally, the DL results by

Table 3.3: Ranked DL results associated with each center reference point.

Vowel Category	Overall DL	x' DL		y' DL		z' DL	
IH	.0073	.0070	(1)	.0120	(1)	.0031	(3)
EH	.0087	.0083	(4)	.0150	(2)	.0025	(1)
AE	.0096	.0073	(2)	.0190	(4)	.0027	(2)
AH	.0097	.0090	(5)	.0170	(3)	.0032	(5)
AA	.0110	.0097	(6)	.0200	(6)	.0032	(4)
ER	.0110	.0080	(3)	.0200	(5)	.0063	(10)
UH	.0120	.0100	(7)	.0210	(7.5)	.0041	(7)
AO	.0130	.0130	(9)	.0210	(7.5)	.0055	(8)
UW	.0140	.0130	(8)	.0240	(9)	.0056	(9)
IY	.0150	.0140	(10)	.0290	(10)	.0037	(6)

axis are also shown for each reference in this table with rankings within that condition shown in parentheses. The rankings by axis agree reasonably well with the overall rankings with the exception of the reversed rankings for [IY] and [ER] for the z' axis. In general, discrimination appears best for the mid front vowels and worst for high vowels with low, central, and retroflex vowels intermediate.

Ambiguous reference points

Although no significant main effect was found for the ambiguous reference points, a rank ordering of the overall DLs and DLs by axis are shown in Table 3.4 for comparison. The

Table 3.4: Ranked DL results associated with each ambiguous reference point.

Vowel Category	Overall DL	x' DL		y' DL		z' DL	
EHAE	.0078	.0058	(1)	.0150	(2)	.0025	(3)
IYIH	.0085	.0094	(5)	.0130	(1)	.0027	(4)
IHEH	.0093	.0064	(3)	.0190	(4)	.0024	(2)
AHAA	.0100	.0095	(6)	.0170	(3)	.0037	(5.5)
AEAH	.0100	.0063	(2)	.0220	(7)	.0023	(1)
AHUH	.0110	.0079	(4)	.0220	(6)	.0037	(5.5)
UHUW	.0130	.0120	(7)	.0220	(5)	.0041	(7)

discrimination trend tends to follow that of the center reference points with better discrimination for references in the front vowel region and poorer discrimination for references in the back vowel region. The range for the overall DLs is somewhat smaller than for the center references. This may be due to the fact that the ambiguous points as a group tend to be more intermediate in terms of articulation and lack the extreme articulatory points which yielded the poorest DLs for the center references. Rank orderings between the axis results were considerably more varied than with the center references.

3.3.3 Analyses by Subject

To evaluate the uniformity of performance across subjects, an ANOVA was computed similar to the first general analysis previously described except subjects were now employed as the first grouping factor and replications served as replicates. A highly significant main effect for subjects was found ($\rho = 0.0006$). A large number of significant two- and three-way interactions were also found, suggesting that subjects were not only significantly different from one another, but that the ways in which they differed varied with the grouping factors.

Because of the varied individual differences between subjects, separate ANOVAs were

computed for each subject based on each subject's own performance across the various grouping factors once again employing replications as replicates. The significance levels for the main effects from these individual analyses are shown in Table 3.5 along with similar results from the overall analysis. Although the significance levels for individual subjects

Table 3.5: Analysis of variance results for effect of conditions overall and by subject.

Variable Grouping	Overall	Subject 1F	Subject 2F	Subject 1M	Subject 2M
Reference Group	ns	ns	ns	ns	***
Axis	**	***	***	*	**
Direction	ns	*	*	ns	ns

*** $p < 0.001$.
** $p < 0.01$.
* $p < 0.05$.
ns: $p > 0.05$.

should be viewed only as an indicator of which subjects strongly followed the direction of the overall results, obvious individual differences are apparent. While the results for three of the four subjects agree with the overall non-significant result for differences between center and ambiguous reference groups, the results for subject 2M (second male) indicate a high level of significance. Similarly, the non-significant results for direction from the two male subjects, 1M and 2M, agree with the overall results, but significant results are found for the female subjects, 1F and 2F.

3.3.4 Analyses by Percentage of Formant Change

As has been discussed previously, $\Delta F/F$ for a fixed distance in *APS* varies with axial orientation, therefore making comparisons of DLs as percentages in Hertz difficult, if evaluated as distances along different axes. To normalize across these differences and enable comparisons between axes, the DLs may be evaluated in terms of percent F change, i.e., ΔF expressed as a percentage of the reference formant frequency value, or $100(\Delta F/F)$, rather than distance.

A normalization of this kind is of particular interest for evaluating differences in DLs related to the x' and y' axes. While the frequency shifts for $F1$, $F2$, and $F3$ related to movement parallel to z' has been shown to be somewhat complicated, frequency shifts for

$F1$ and $F2$ related to movement parallel to x' and y' is similar. Changes parallel to either axis result in approximately equal shifts for both formants when expressed as percentages of F change with the difference between axes now being the pattern of spectral change. The differences between DLs related to direction along these two axes can be further normalized by utilizing the absolute values (i.e., disregarding the signs) for the percentages of F change.

To evaluate the difference between DLs associated with the two spectral patterns induced by movement parallel to the x' and y' axes, a repeated-measures ANOVA was computed employing the four subjects as replicates. The analysis was similar to that used for the first general analysis except that the absolute values of percent $F2$ change were used as the DL variable instead of log unit distance. Since percentages calculated from $F1$ change and $F2$ change are approximately equal for any given movement along each axis, these analyses could have been computed using percentages of $F1$ change which should have yielded the same results. A 2 x 2 x 2 x 2 factorial design was used for the analysis, comparing the two reference point groups (center vs. ambiguous), the two axes (x' vs. y'), the two axial directions (positive vs. negative axis), and the two replications.

The results of this analysis proved to be very similar to the results of the first general analysis discussed in section 3.3.1. There was no significant main effect once again for differences between center and ambiguous reference point groups ($\rho = .097$) or differences between positive and negative directions along continua. However, a significant main effect was found for differences between DLs associated with the x' and y' axes ($\rho = .033$) and for differences between replications ($\rho = .043$). A significant interaction between the axis and direction factors ($\rho = .034$) was also noted and follows the same pattern as was discussed in section 3.3.1. The significant effect of most interest from this analysis however, is the difference between x' and y' axes. The average percent $F2$ change for continua parallel to the x' axis was 1.47% and to the y' axis, 1.81%. This result suggests that there is a significant difference in how subjects discriminate the two different spectral patterns represented by movement parallel to these axes. Subjects seem to better discriminate patterns where $F1$ and $F2$ move in like directions (x') than patterns where $F1$ and $F2$ move in opposing directions (y').

Separate ANOVAs were once again computed to evaluate differences within the two

reference point groups using the percent $F2$ change DLs. The same factorial design was used for the analyses as was described above, except that the reference point factor now had 10 levels for the center reference point analysis and seven levels for the ambiguous reference point analysis. The results of these analyses are compared to the overall analysis in Table 3.6. The patterns of significant effects for the two analyses are quite disperate.

Table 3.6: Significances of factors from overall and individual reference group analyses-of-variance of DLs expressed as percent F2 change.

Factor	Overall	*center* references	*ambiguous* references
Reference Group	ns	$\rho = .0008$	ns
Axis	$\rho = .033$	ns	$\rho = .030$
Direction	ns	$\rho = .046$	ns
Replication	$\rho = .042$	ns	ns

As before, a significant main effect was found for differences between the ten center reference points, but no similar effect was found for differences between the seven ambiguous reference points. The axis factor was significant for the ambiguous reference group analysis, but not for the center reference group analysis. The opposite was true for the direction factor with a significant result found for the center reference point analysis, but not the ambiguous reference point analysis. Each analyses had one significant two-way interaction.

For the center references analysis the reference-by-direction interaction was significant ($\rho = .002$). This interaction may be due to considerable differences in discrimination by axial direction between references, although overall discrimination is significantly better for positive axial directions. For the ambiguous references analysis the reference-by-axis interaction was significant ($\rho = .011$). A similar reasoning can be applied here, that is, while overall discrimination is significantly better for the x' axis, this pattern is reversed for several of the individual reference points.

In summary, we find that, when the differences in DLs for continua parallel to the x' and y' axes are expressed as absolute changes in formant frequency relative to the reference, significant statistical factors remain basically the same as was found for the differences in

DLs expressed as distance. There is no significant difference between the center and ambiguous reference groups. Within these groups, however, the center references are significantly different from one another, but not the ambiguous references. Direction along axes was not found to be a significant factor overall, although significant differences exist for the center references with the better discrimination found for positive axial directions. Replication was a significant factor overall, but not at the individual reference group level. Perhaps of greatest interest, however, is that a significant difference was still found for discrimination between continua associated with the x' and y' axes. This difference is largely accounted for by significant differences for this factor within the ambiguous references, although the same pattern of difference, that is, continua associated with the x' axis are better discriminated than continua associated with the y' axis, is seen for the center references as well. The significant difference found between these axes also implies that there is a difference in how the two types of spectral pattern change associated with these axes, that is, parallel formant movement vs. opposing formant movement, are discriminated and that the better discrimination is for formant patterns exhibiting parallel movement.

3.3.5 Analyses of Movement in z'

As was stated earlier, movement along continua parallel to the z' axis represents unequal percent changes in $F1$, $F2$, and $F3$ making these continua less comparable to movement parallel to x' or y'. Therefore, three separate analyses have been computed on the log unit distance measures for DLs in z'. These analyses follow the same format as those for the x' and y' comparisons where the first analysis compares between center and ambiguous reference point groups, and the following analyses compare within the two groups.

The first analysis was a 2 x 2 x 2 factorial design ANOVA utilizing subjects as replicates and comparing center vs. ambiguous reference points, positive vs. negative directions along the axis, and the two replications. The within-group analyses followed the same design except that the first factor had ten levels for the center reference group and seven levels for the ambiguous reference group. The results of these analyses are shown in Table 3.7 and indicate that significant differences exist not only between center and ambiguous reference point groups, but also within each group.

Table 3.7: Analysis-of-variance results for effect of conditions overall and by reference group for z' continua.

Variable Grouping	Overall	*center* references	*ambiguous* references
Reference Group	$\rho = .024$	$\rho = .0008$	$\rho = .011$
Direction	ns	ns	$\rho = .025$
Replication	ns	ns	ns

For the between-group analysis, the ambiguous references as a group were discriminated significantly better than the center references. No significant differences were found for direction or replication. Significant interactions were found for group-by-direction ($\rho = .048$) and group-by-replication ($\rho = .024$). The group-by-direction interaction may be the result of significantly different discrimination for direction found among the ambiguous references interacting with virtually no difference in discrimination for direction among the center references group. The group-by-replication interaction may reflect the product of opposing trends for replications between the two groups.

The results of the within-group analyses are similar to the between-group analysis except that the direction factor is significant for the ambiguous reference continua with negative-going continua the better discriminated. No significant interactions were found for either of these analyses. The rank ordering of DLs by reference along the z' axis were shown previously in Tables 3.3 and 3.4.

3.3.6 Overall Discrimination by Reference

Examination and discussion (See Section 3.3.2) of Tables 3.3 and 3.4 indicated that the DLs associated with individual references can vary considerably for any of the three axes. Individual formant locations and patterns of $F1$, $F2$, and $F3$ may be considered as a plausible explanation for differences in the DLs in general between the various reference points. The average percent $F2$ change ($100(\Delta F2/F2)$ for all x' continua by reference point are shown in Figure 3-7 along with the frequencies of $F1$, $F2$, and $F3$ of the reference points. A similar illustration for all y' continua is shown in Figure 3-8. Values for like

Figure 3-7: Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for x' continua.

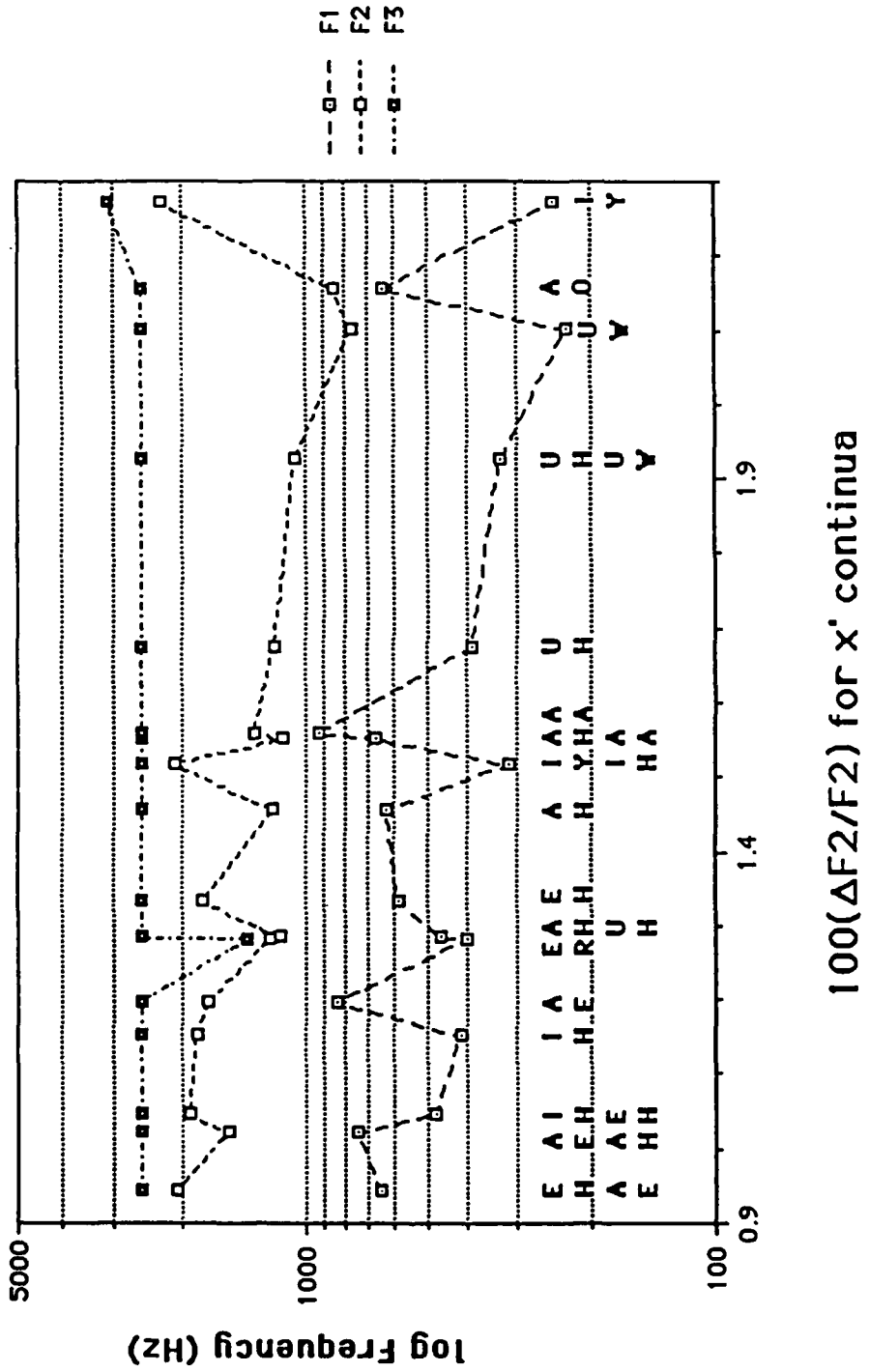
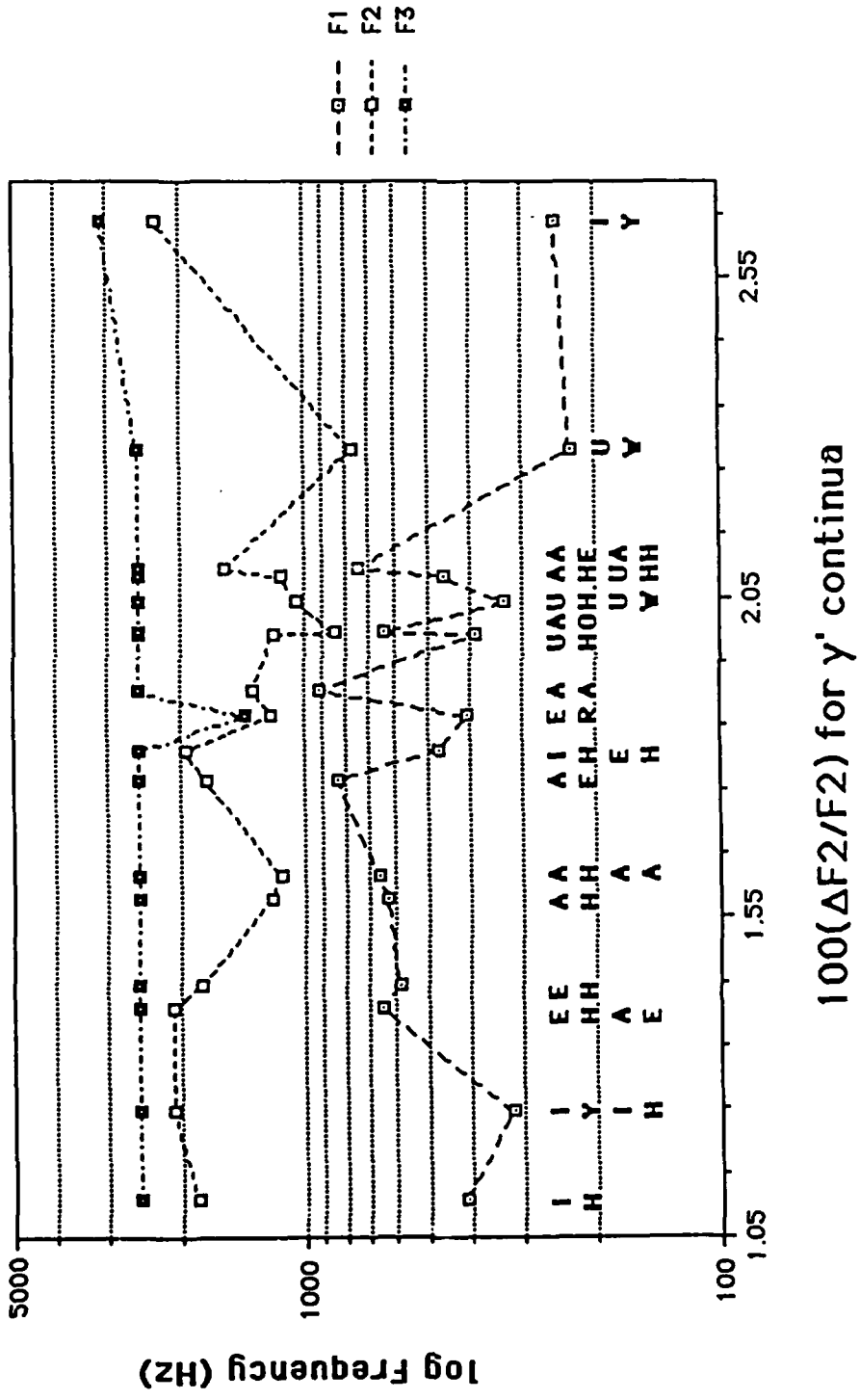


Figure 3-8: Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for y' continua.



formant positions are connected by broken lines for ease of illustration. For the x' continua in Figure 3-7 we find a general trend for smaller DLs to be associated with higher values of $F2$, and with DLs increasing with decreasing values of $F2$. A pattern associated with $F1$ is less discernible but, in general, given a value of $F2$, lower values of $F1$ are associated with smaller DLs and higher values of $F1$ with larger DLs. Exceptions to both of these patterns are DLs associated with the /IYIH/ and /IY/ reference points. For these cases, we must assume that the extreme distances between $F1$ and $F2$ reduce discriminability.

Examination of Figure 3-8 for the y' continua reveals a somewhat similar pattern. The formant patterns for reference points associated with the best and worst DLs seem to generally follow the trends noted for the x' continua, but the majority of reference points lie intermediate to these and exhibit considerable variation in formant pattern.

The ordering of reference point formant patterns for z' continua, as seen in Figure 3-9, exhibits the same trend for decreasing $F2$ with increasing DL as seen for the x' and y' continua. However, the trend in the $F1$ pattern seems to be reversed from the $F1$ pattern seen in the previous figures, with $F1$ now decreasing with increasing DL, paralleling the $F2$ trend. Notable exceptions are found for /IY/, which is better discriminated along this axis, and /ER/, which, presumably due to the close proximity of $F2$ and $F3$, is the most poorly discriminated reference.

3.3.7 Single vs. Multiple Formant Movement

In addition to the continua previously specified which reflected multiple simultaneous formant changes, eight other continua reflecting single formant changes were also evaluated twice by three of the four subjects. These continua (see figure 3-10) represent straight lines in *APS* which emanate at 60 (labeled continuum 2), 120 (continuum 4), 240 (continuum 8), and 300 (continuum 10) degrees relative to the x' axis from the [EH-AE] and [AH-UH] ambiguous reference points. Points along the 60° and 240° continua represent vowel tokens where only $F1$ varies relative to the formant values of the reference point. Points along the 120° and 300° continua represent vowel tokens where only $F2$ varies relative to the formant values of the reference point. Points located 0.02 log units away from the reference point along these continua represent changes of approximately ± 3.3% in $F1$ or $F2$ relative to

Figure 3-9: Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for z' continua.

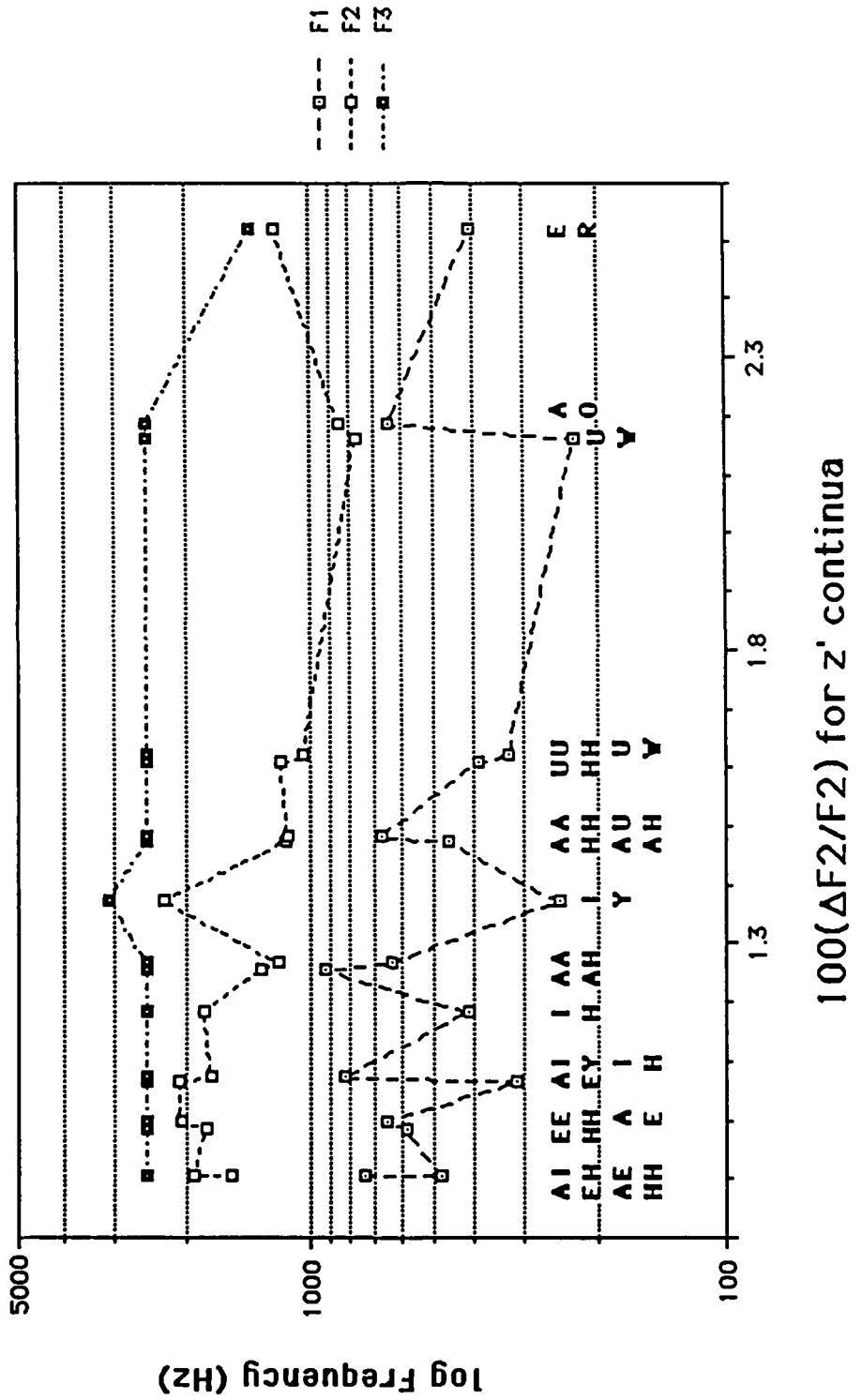
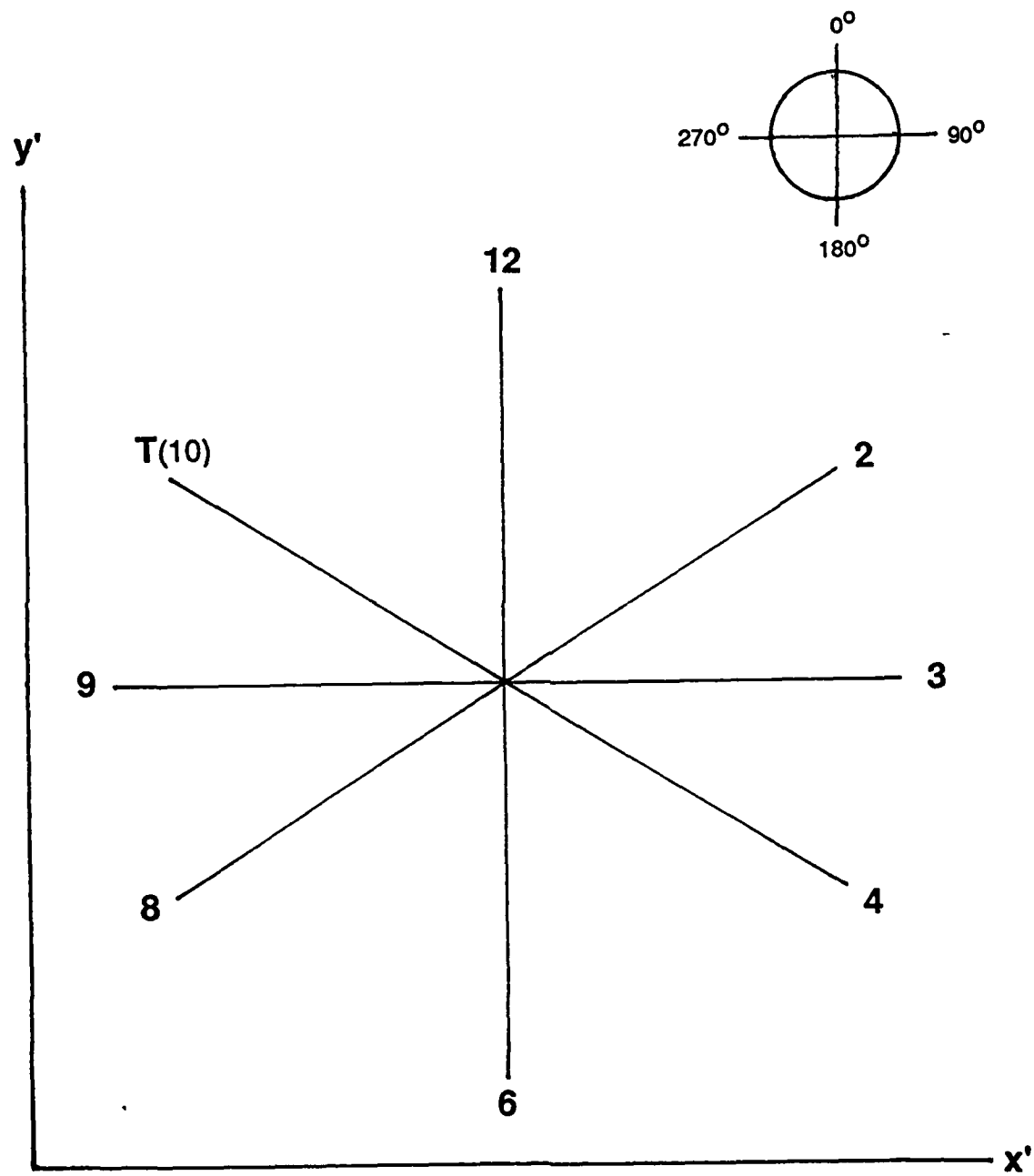


Figure 3-10: Locations in APS $x'y'$ coordinates of single-formant-change continua relative to multiple-formant-change continua.



the reference point formant frequencies, similar to the percentage of formant change found for points at this distance along continua parallel to the x' axis. No evaluation was made for changes in $F3$ alone.

A single-factor ANOVA indicated that differences in percentage of formant change between $F1$ and $F2$ were not significant. Therefore, the DL estimates along continua exhibiting single formant changes were directly compared to the DL estimates from continua parallel to the x' and y' axes exhibiting multiple formant changes emanating from the same reference points. Several questions of interest may be addressed by these comparisons. First, are DLs for single-formant-change continua significantly different from DLs for multiple-formant-change continua? Secondly, are DLs for single-formant-change continua more like the DLs for continua parallel to x' or y'?

The comparisons were once again based on DLs expressed as absolute percentages of $F1$ and $F2$ change as was done previously for comparison of x' and y' continua. The first comparison was a 2 x 2 x 2 x 2 x 2 factorial ANOVA utilizing the three subjects as replicates. The factorial design compared by group (i.e., single- vs. multiple-formant-change), by reference (AHUH vs. EHAE), by continua (continuua 2, 4, 8, and 10 vs. continua 3, 6, 9, and 12), by direction (continua 2, 4, 3, and 12 vs. continua 8, 10, 6, and 9), and by replication. The results of this analysis indicated only one significant main effect which was for the by-group factor ($\rho = .023$). This effect indicated that the DLs for the multiple-formant-change continua were significantly smaller ($\overline{x} = 1.42\%$) than the DLs for the single-formant-change continua ($\overline{x} = 1.89\%$). No significant interactions were found. This analysis was repeated with the groupings for the by-continua factor reversed such that continua 4 and 10 were grouped with the y' continua and continua 2 and 8 were grouped with the x' continua. This analysis yielded virtually the same results. Thus the results of these analyses suggest that the answer to the first question is yes, DLs for single-formant-change continua are significantly different from DLs for multiple-formant-change continua.

Recall that movements along the x' and y' axes represent two distinctly different changes in spectral pattern and that significant differences exist between DLs for these two axes. Although we find that DLs for single-formant-change continua are significantly different from the DLs for the x' and y' axes when grouped together, it is of interest to know whether

discrimination of changes in spectral patterns associated with varying a single formant is different from discrimination of changes in the two spectral patterns associated with the x' or y' axis (i.e., parallel or opposing shifts in $F1$ and $F2$) or more similar to one of them. To address this question, analyses of DL values expressed as absolute percentages of $F1$ or $F2$ change for the single-formant-change continua were compared first with corresponding values from continua parallel to the x' axis and then similarly with the y' axis. These comparisons were by way of ANOVAs using the same factorial design as for the first analysis and the same groupings for the first two factors (i.e., single- vs. multiple-formant-change and AHUH vs. EHAE). However the third grouping factor now becomes $F1$ vs. $F2$ which groups DLs for continua 4 and 10 with x' or y' $F1$ DLs and continua 2 and 8 with x' or y' $F2$ DLs. Furthermore, grouping for the direction factor was now by positive or negative formant change and not by direction in the space.

These analyses indicated that the DLs for single-formant-change continua are significantly different from the x' DLs ($\rho = .017$) but not significantly different from the y' DLs ($\rho = .477$). The DLs for the x' axis are significantly smaller (($\bar{x} = 1.03\%$) than the DLs for either the y' axis ($\bar{x} = 1.81\%$) or the single-formant-change continua ($\bar{x} = 1.89\%$). This result suggests that the changes in spectral pattern being discriminated along single-formant-change continua are more like the spectral pattern changes associated with y' continua than with x' continua.

3.4 Summarization and Discussion of Experiment II

In summary, we find that, based on the results of this experiment, the average DLs for the x', y', and z' axes are approximately .01, .02, and .004 log units respectively. As applied to the pilot mapping of the $z' = 0.70$ plane in .01 log unit steps presented at the end of Chapter 2, these results suggest that such a grid size should reflect sufficient resolution to estimate vowel category boundaries accurately along x' and y' and represents a good compromise in resolution if a single grid size is dictated. However, we can expect redundant boundary information concerning tokens along y', and, had additional planes been mapped at .01 log unit resolution, insufficient boundary information pertaining to tokens along z'.

If a variable grid size were to be used for mapping, the number of tokens required could

be cut in half if the grid size along y' were increased to .02 log units. The gain in efficiency brought about by this token reduction would be lost however, if planes of tokens along z' must be mapped at .004 log units. A reasonable alternative may be to consider mapping only the z' area associated with the perception of retroflection since it has been shown in Experiment I and elsewhere that changes in $F3$ affect the perception of non-retroflex vowels very little, at least for American English.

Results from this experiment did not reveal significant differences in discrimination between reference points located within phonetic vowel categories and reference points located at phonetic boundaries. Investigation of the possible differences in perceptual discrimination of synthetic vowel sounds from within and between phonetic categories has been active since 1962, when Fry and colleagues concluded that there were no differences in discrimination for the case of isolated synthetic vowel sounds. Since that time a number of investigators have attempted to detail more specifically how and why vowel (and consonant) discrimination may vary relative to phonetic classification. Among the variables which have been investigated are isolated vowels vs. consonant-bound vowels (Stevens, 1968; Sachs, 1969; Mermelstein, 1978), duration (i.e., short vs. long) (Sachs, 1969; Pisoni, 1973; Mermelstein, 1978), and psychophysical methods (Sachs, 1969; Pisoni, 1973; Macmillan, et al., 1986; Macmillan, Goldberg, and Braida, 1988). Taken as a whole, this past research would suggest that differences in discrimination related to phonetic boundary proximity are most likely to occur with synthetic vowels of short duration, vowels in a consonant context, and paradigms utilizing a roving ABX or same/different (AX) task. The present experiment found no such differences and, given that the experiment utilized synthetic vowels of reasonably long duration in an isolated context and were measured with an adaptive, cued-2IFC task, we should not expect to find significant differences in discrimination related to phonetic boundary proximity.

A more perplexing issue pertains to the significantly smaller DLs found in this experiment as opposed to similar experiments in the past. In general, the DL for changes in $F1$ or $F2$ when ΔF is expressed as a percentage of the formant frequency has consistently been found to be on the order of 3% to 5% (Flanagan, 1955; Kakusho and Kato, 1968; Mermelstein, 1978; Nord and Sventelius, 1979). For multiple simultaneous formant changes, this

experiment found average DLs to be approximately half that, on the order of 1.5% to 1.8% of the formant frequency. This result in and of itself would be acceptable, if varying multiple formant frequencies simultaneously produces additive information which in turn induces an increase in discrimination. Mermelstein (1978) proposed a weighted additive model based on $\Delta F1$ and $\Delta F1$ for single formant variation to predict the increased discrimination he reported for simultaneous variation of $F1$ and $F2$. Additionally, Carlson, Granström, and Klatt (1979) found slightly increased perceptual distances for simultaneous variation of $F1$ and $F2$ relative to perceptual distances for single formant variation. The difficulty here lies in the fact that DLs found in this experiment for single formant variation, while shown to be significantly different from the multiple-formant-variation DLs, are still on the order of 2%, definitely less than results from past research would suggest.

A number of factors may be considered as potential explanations for the differences in DLs from the present experiment and similar work in the past. All of these factors are related to the psychophysical methods and experimental protocols employed. The first consideration is subject training. Neither Flanagan (1955) nor Nord and Sventelius (1979) mention any training for their subjects and Mermelstein (1978) reports that subjects listened to about ten pairs of stimuli as familiarization prior to testing. In the present experiment, each subject was trained with a minimum of four runs on 14 different continua, about 3400 trials, prior to testing. Despite this degree of training, subjects still demonstrated significant improvement in their discrimination performance with replication. Although Flanagan (1955) utilized five repetitions and four subjects and Nord and Sventelius (1979) an average of 39 repetitions and 27 subjects, only Mermelstein (1978) makes mention of any significant differences between subjects. It is most probably safe to assume that subjects were better trained in the present experiment than in the studies mentioned above and this difference could yield more sensitive discrimination results.

Next, we consider the differences in the experimental methods used to estimate DLs as a possible factor for differences in discrimination. All three experiments cited from the past utilize a two-interval, same/different (2IAX) task from which the percentage correctly identified as "different" is plotted and the 50% correct point estimated. Although not explicitly stated in all three vowel DL studies cited, this task is often designed in a roving

discrimination pattern, i.e., any two stimuli among the references and their variations may be paired randomly, and results are corrected for guessing based on assumptions underlying "low-threshold" theory interpretations of discrimination experiments proposed by Luce (1963). The present experiment, on the other hand, utilizes a two-interval, forced-choice (2IFC) task with a cued, fixed standard and is interpreted in terms of signal detection models. Macmillan, Kaplan, and Creelman (1977) and Macmillan, Goldberg, and Braida (1988) both investigated the implications of these two methods on discrimination in speech research. Both concur that with the 2IAX task, response bias tends to be large, listening strategies non-optimal, and performance substantially lower than performance for yes/no (YN) or 2IFC tasks. The 2IFC task, on the other hand, is relatively free of response bias and a more optimal method of estimating true sensitivity. Additionally, Macmillan, et al. (1988) demonstrated that fixed discrimination, like that used in the present experiment, is more sensitive than roving discrimination, most likely to due to decreased memory requirements and/or lower stimulus uncertainty. The inter-stimulus interval (ISI) also plays a larger role in roving discrimination designs (Pisoni, 1973). Pisoni found that a .5 sec ISI (the past DL experiments cited used .4-.5 sec ISIs) yielded somewhat lower discrimination sensitivity than the .25 sec ISI (used in the present experiment) for synthetic vowels of longer (300 msec) duration.

The present experiment employs a cue or standard (the reference token) preceding each observation interval. Khazatsky (1985), in modeling sensitivity changes, demonstrated that standards can improve performance, particularly in the region of the standard, for identification tasks. The use of trial-by-trial cues in discrimination tasks is somewhat less clear, although cues have been found to yield improved performance (Greenberg, 1962). The general thinking is that cues or standards help reduce subjects' uncertainty and reduce the influences of internal references. In overview, Robinson and Watson (1972) state that "... performance can be improved by providing the listener with as much information as possible about the to-be-detected signal." (p. 111).

The last methodological issue to be raised is the use of adaptive and non-adaptive testing paradigms. Kollmeier, Gilkey, and Sieben (1988) evaluated several adaptive staircase rules with several psychophysical procedures (2AFC and 3AFC) and found that while modelings

of the various tasks suggest that results from adaptive and non-adaptive techniques should be equivalent, human data indicated that the adaptive techniques tended to yield lower thresholds. Although empirical evidence is scarce, it is also generally held that subjects are able to learn a task more rapidly when adaptive methods are employed.

Taken together, the differences in methodology discussed here may well explain the differences in the results between the present experiment and similar past work. Subjects in the present experiment were probably better trained and numerous considerations were given to the experimental protocol to minimize uncertainty and maximize discrimination sensitivity.

A question of interest arises in consideration of the differences between formant DLs for single- and multiple-formant variation. Are the DLs related, that is, can multiple-formant DLs be predicted from single-formant DLs? The model proposed by Mermelstein (1978) considers changes in $F1$ and $F2$ to independently contribute information to discriminability of stimuli where both formants are simultaneously varied. In the model, the composite ΔF for these stimuli is calculated by

$$(\omega_1(\Delta F1)^2 + \omega_2(\Delta F2)^2)^{1/2} \tag{3.2}$$

where ω_1 and ω_2 are unspecified weighting factors. Mermelstein implies that the weighting factors are related to the relation of $\Delta F1$ and $\Delta F2$, however, our own attempts to replicate his predictions from his DL results have failed. If the weighting factors should be related to the percent change of each formant, then equal weightings of 1 may be used to predict the present data. The resulting predictions for the average multiple-formant-change DLs from the average DLs for single-formant change are quite accurate, within 2 Hz, of the actual average results. Thus it would appear that $F1$ and $F2$ may equally contribute information for discrimination. If the model were to be expanded to include $F3$, with $F3$ also contributing information on an equal basis, we could use the equation

$$(\omega_1(\Delta F1)^3 + \omega_2(\Delta F2)^3 + \omega_3(\Delta F3)^3)^{1/3} \tag{3.3}$$

to estimate changes for continua associated with z' in the present experiment. Since a DL for single-formant variation in $F3$ has not been estimated presently or in the past literature, we shall, for the moment, assume the prediction equation is accurate and solve for the missing

single-formant-variation DL value for $F3$. If weightings are assigned to reflect the relative percent formant changes resulting from movement parallel to z' ($\omega_1 = .337$, $\omega_2 = .663$, and $\omega_3 = 1$), the resulting DL for single-formant variation in $F3$ should be about 1.87% of the reference $F3$. While, intuitively this value seems small, it is in line with the single-formant-variation DLs found for $F1$ and $F2$, and suggests that such a model may have merit.

Flanagan (1955) found that DLs for single formant manipulation were affected by the relative proximity of two neighboring formants. He reported asymmetries in the positive and negative frequency DLs, particularly for $F2$, with some formant combinations. In general, DLs decrease for the direction toward a neighboring formant in close proximity and increase for the direction away from the adjacent formant. Flanagan suggested that these asymmetries were due to larger increases in formant amplitude with closely neighboring formants than with formants exhibiting larger spacing in frequency. Mermelstein (1978) found a similar pattern in his results. Nord and Sventelius (1979) likewise found asymmetries in their replication of Flanagan's work, but in the opposite direction, and were forced to reject Flanagan's conclusions regarding formant intensity relations. They noted that for small shifts in $F2$ (i.e., < 50 Hz), there was no drastic change in level for the $F2 - F3$ complex and suggested that perhaps the increase in intensity build-up will not appear unless the proper auditory analysis is made. Additionally, they found good correlation between their discrimination curves and spectral distance measures (described by Plomp, 1970) based on 1/3 octave band filter analysis.

For the present experiment, significant differences between axial directions were found only among center references for the x' and y' axes and among ambiguous references for the z' axis. In general, positive axis directions yielded better discrimination for the x' and y' axes and the negative direction yielded better discrimination for the z' axis. These results are difficult to compare directly to past results noted above in that positive and negative frequency DLs are confounded for continua parallel to y', where $F1$ and $F2$ move in opposing directions. Despite this difficulty, closer examination of the results indicates no specific pattern related to direction of formant frequency change and formant proximity. The lack of differences related to direction of formant frequency change in the present

experiment may, once again, be attributable to the differences in methodology previously discussed.

Section 1.2.6 outlined general trends seen in formant relations potentially related to differences in discrimination for continua associated with the various reference points. A predominant trend was that discrimination ability seemed to decrease with decreases in $F2$. While patterns in $F1$ related to discrimination ability were more variable with axis and more difficult to discern, it is clear that $F1$ also plays some role in the general discriminability of the various reference points.

Another possible avenue of explanation for discrimination differences between references is through analysis of their auditory spatial-frequency patterns. Spatial frequency, as it relates to vowels, considers the relation between two adjacent peaks in the spectrum as one modulation of a frequency expressed in cycles/octave. Likewise, the spatial frequency measured between $F0$ (or a reference) and $F1$ can be thought of as a relative location measure for the spectrum. Although an extensive analysis regarding spatial frequency will not be undertaken here, a preliminary investigation suggests that these measures may be great interest.

To further explore the possible accountability of discrimination differences between reference points related to the three spectral patterns mentioned previously, linear multiple regressions were calculated. Three sets of independent variables were used. The first set contained the frequencies for $F1$, $F2$, and $F3$ of the seventeen reference points. The second set contained the the x, y, and z values for the reference points. Recall that these values are log ratios of formants, where $x = log(F3/F2)$, $y = log(F1/SR)$, and $z = log(F2/F1)$. The third set contained the reciprocals of the x, y, and z values which are related to the spatial frequencies of the formant relations. The dependent variables for these analyses were the DL results expressed as percentage change in $F2$ (i.e., $100(\Delta F2/F2)$ for continua associated with the x', y', and z' axes. The results of these analyses are shown in Table 3.8.

For the DLs associated with x', all three variable sets are able to account for a significant proportion of the variance. However, the reciprocals of the log ratios of formants, related to spatial frequency, account for the greatest proportion and are highly significant ($\rho = .0001$).

Table 3.8: R^2 values from multiple regression analyses of DL results and specified variable sets (See text).

Variable Set	DL x'	DL y'	DL z'
Formants	.565	.248	.744
Log ratios	.605	.256	.832
Reciprocals	.788	.390	.467

Within the independent variable coefficients, the reciprocal of $log(F1/R)$ is the most highly significant ($\rho = .0001$), but the reciprocal of $log(F2/F1)$ is also significant ($\rho = .003$). This finding suggests that differences between references for these continua may be potentially accounted for primarily by the the relative location of $F1$ and secondarily by the relation of $F2$ to $F1$.

For the DLs associated with y', none of the three independent variable sets are able to significantly account for the variance. However, of the three sets, the reciprocals come the closest ($\rho = .084$). Once again, the ordering of independent variable coefficients indicates the reciprocal of $log(F1/R)$ accounts for largest amount of variance alone. The results suggest that none of formant-related variables considered can adequately account for DL differences in references for the y' continua.

For the DLs associated with z', all three independent variable sets are able to significantly account for the variance at the $p < .05$ level, however, the simple formant ratios provide the most significant accounting ($\rho = .0001$). The ordering of accountability within the independent variables remains as seen for the other axes, i.e., $log(F1/R)$, $log(F2/F1)$, and $log(F3/F2)$.

These analyses, while by no means exhaustive, provide evidence that spatial frequency relations may well play a part in the differences between DLs related to the reference formant patterns and should be further explored. However, we also continue to find, at this point, unexplainable differences related to the axis of movement.

Recall that three distinct patterns of spectral change are associated with movement away from reference points and that each pattern is related to one of the x', y', or z' axes.

Movement parallel to x' results in equal-percentage parallel shifts in $F1$ and $F2$ and may be thought of as reflecting Fant's (1960) distinction for grave/acute. The spectral patterns generated along these continua maintain a constant spatial frequency relation between $F1$ and $F2$, but vary the spectral location of this frequency. Movement parallel to y' results in equal-percentage opposing shifts in $F1$ and $F2$ and may be thought of as reflecting Fant's distinction for compact/diffuse. Spectral patterns generated along these continua vary the spatial frequency relation between $F1$ and $F2$, but maintain the spectral location of these frequencies. Movement parallel to z' results in unequal-percentage parallel shifts in $F1$, $F2$, and $F3$ and, while similar to x' movement, results in a more total shift of the spectrum. The spectral patterns generated along these continua maintain a constant spatial frequency relation between $F1$ and $F2$ and $F3$, but vary the spectral location of these frequencies.

Previous analyses have indicated that there is a small, but significant difference in discrimination between the spectral patterns associated with the x' and y' axes. We must assume for the present time that these differences are related to the differences in spatial frequency relations and spectral location noted above. While discrimination patterns of z' continua are somewhat similar to patterns seen for x' continua, the consideration of changes in $F3$ results in a distinct third pattern of discrimination requiring its own explanation. It should be the goal of future reseach to further elucidate the processes from which these differences in discrimination emanate.

Chapter 4

Final Comments and Implications for Future Research

It is suggested that the work presented in this dissertation be considered only a small example of the basic research yet required to understand the processes of speech perception. In addition, this work should also be considered exploratory in nature and, to that extent, the results preliminary. The results of the experiments presented here will hopefully open more doors to possible further investigation than they close as being definitive statements of fact. And that, I think, is as it should be.

Although the finding that abutting and non-overlapping target zones can be constructed based on synthetic vowel sounds lends validity to the target zone concept and the utilization of such zones in theories of speech perception, many issues remain unresolved and untested. How best to represent target zones, their boundaries, and the speech signal itself in order to accurately model the processes of human vowel perception will require considerable amounts of continued investigation. An immediate need for further research which has already been called for is the mapping of zone boundary areas at higher resoluton. Different methodological approaches, however, should be investigated to determine how judgements of tokens near category boundaries are influenced by subject task and other procedural variables. In addition, many simple extensions of Experiment I are also immediately implied. Similar mapping experiments with other fundamental frequency ($F0$) contours should im-

pact on current strategies for talker normalization and potentially related changes in zone boundaries. Considerable attention has been given to the potential differences in vowel perception between vowels in a consonantal context and vowels spoken in isolation (See Strange, 1989 and Neary, 1989 for discussions of these issues). Additional insights may be gained into these issues by comparing mappings of synthetic versions of vowels in and out of consonantal context.

Another dimension for continued investigation relates to the issue of non-vowel sounds located in the vowel space. As was discussed in the introduction to Chapter 2, portions of the unaccounted-for vowel space in target zone estimations based on natural speech productions (figures 1-9 and 1-10) may represent speech sounds other than vowels of American English or sounds not representative of speech at all. Miller and Hawks (1986) presented data which suggested that the initial portions of oral consonants and the voiced portions of fricatives may be best represented in areas of the space adjacent to vowel target zones. Additionally, it has been discussed that some subjects reported certain tokens were best identified as examples of front rounded vowels used in languages other than English. Since no effort was made in Experiment I to differentiate American English vowel sounds from other vowel or non-vowel sounds, some of the area within the synthetic-speech-based target zones may be misrepresented. Further research is required to resolve this issue.

The target zone concept within the *APS* format lends itself well for investigating aspects of perception and production for vowels of all languages. Some preliminary studies have already considered target zones for production data from Greek and German (Jongman, Fourakis, and Sereno, 1989) and perceptual data from one speaker of Greek (Fourakis and Hawks, 1990). In addition to comparative cross-language studies, these approaches may also prove valuable for investigating the effects of bilingualism and second-langauge acquisition on perception and production.

The utilization of approaches like those demonstrated in Experiment I for mapping perceptual responses need not be limited to vowels. Similar mapping studies for virtually any class of speech sounds is certainly plausible. Along with this plausibility comes all the possible variations already discussed for vowels. Consideration of these possible extensions suggests that the work presented here may well serve as only a small foundation on which

to build a sizeable body of new knowledge in speech science.

Of the number of findings related to the estimation of difference limens for multiple simultaneous formant changes of vowel-like sounds reported in Experiment II, perhaps the most important is that the *APS* vowel space, at least in terms of axis-parallel movement in the x', y', z' coordinate system, is not represented in equal perceptual dimensions. This finding impacts not only on considerations for target zone estimation and zone boundary representation, but also creates difficulties in implying any intuitive sense about the movement of the speech signal we are visualizing in these transformed dimensions. An acceptable condition would be to find that fixed distance movements parallel to the original x, y, z axes in *APS* do result in equal DLs, however, this too is probably not the case. Although percentages of formant change for fixed distance movements parallel to any of the x, y, z axes are equivalent, which formant or formants are allowed to vary is still axis dependent. Movement parallel to the x axis results in only $F3$ variation, while similar movement to y yields variations in all of the first three formants, and for z, variations in $F2$ and $F3$. Thus, no coordinate system currently utilized in the context of *APS* can be considered reflective of dimensions scaled in equal perceptual units for discrimination sensitivity. This was also the conclusion of Macmillan, et al. (1988), based on their attempts to account for differences in discrimination sensitivity along a vowel continuum through the examination of distances between the locations of continuum tokens in the *APS*.

Further investigative attention must be given to the implications of unequal perceptual scaling in *APS* in terms of how it may impact on practical and theoretical aspects of modeling human speech perception. Should equal perceptual scaling be required, an alternative coordinate system or further transformations of existing systems will be implied. Figure 4-1 demonstrates one possibility of a simple transformation. In this figure the synthetic-speech-based target zones for one z' plane are shown with the y' axis "warped" to reflect dimensions that are perceptually more equivalent.

A question of some import pertaining to perceptual considerations of the *APS* is what inferences can justifiably be drawn from the results of Experiment II which are reflective of human speech processing? The methodological considerations given to Experiment II were intended to force subjects into non-phonetic auditory processing modes, minimize the

possible effects of internal standards, and provide a relatively low-uncertainty environment for discriminating differences between two complex sounds. The issue can certainly be raised that such an experiment most probably does not reflect the processes used by listeners in everyday speech communication, and thus becomes another example of what humans "*can* do" and not what they "*do* do." This is, however, an issue which is ever present in speech research as well as other areas of science dealing with human abilities and capabilities and can never be completely dismissed. Many studies have been undertaken attempting to quantify the perceptual aspects of complex sound discrimination in both phonetic and non-phonetic contexts. The results of these studies vary as widely as the questions posed and the methodological procedures employed in them. Given that the true dimensions of the processes used in everyday speech perception have yet to be defined, the results of Experiment II potentially reflect the opposite end of the discrimination continuum. That is, the experiment attempted to demonstrate the maximum perceptual discrimination of vowel-like sounds resulting from a non-phonetic auditory processing mode within the *APS* context. If the results are representative of this description, we may at least rest assured that no smaller perceptual unit need be considered and that discrimination ability of vowels in natural speech communication modes should be based in larger perceptual units. The quantification of these units however, must await future research.

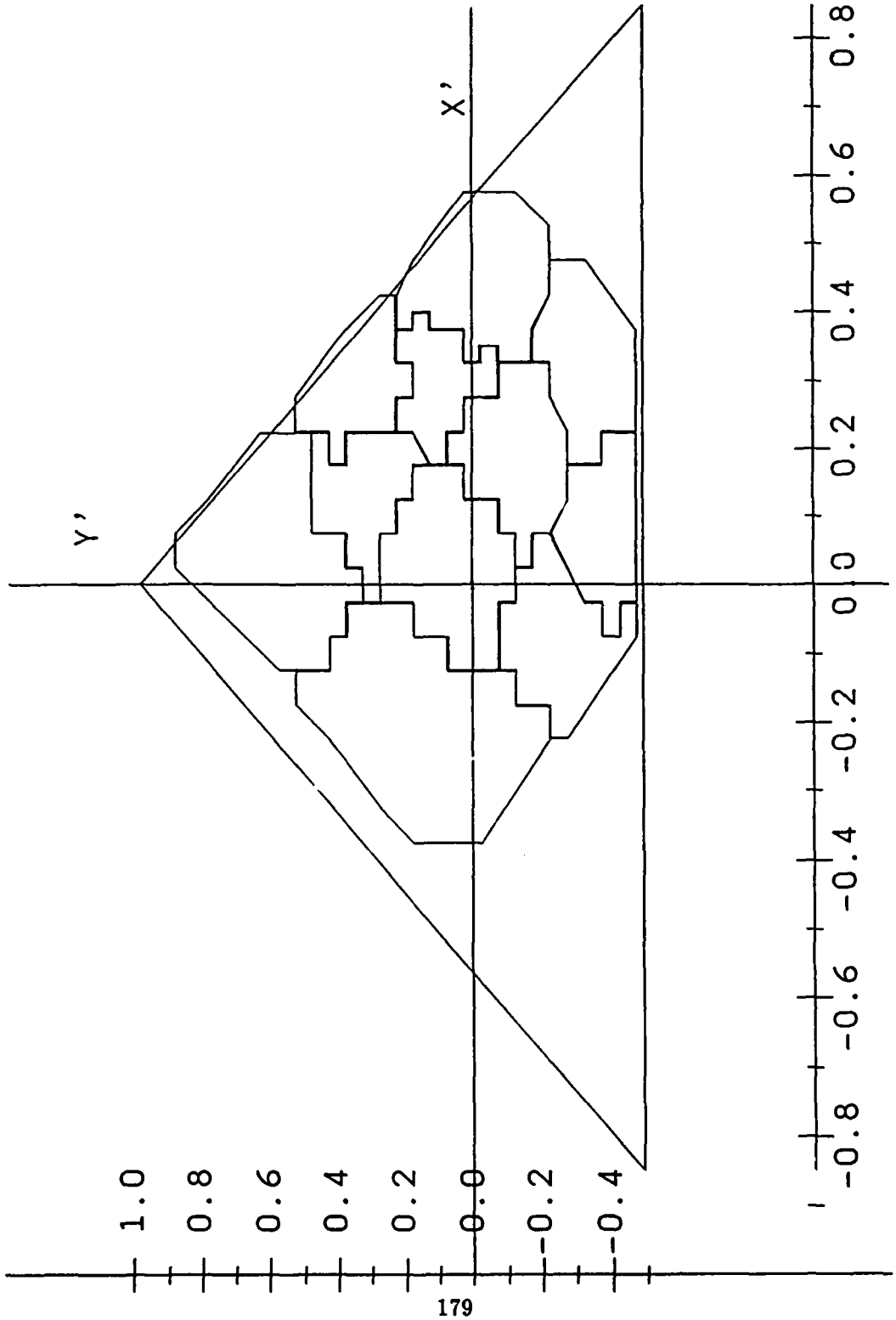
Additional motivation for further research can be spurred by other findings resulting from Experiment II. While the results found for discrimination of vowel-like sounds with simultaneously varying, multiple-formant changes may be generally accounted for by a relatively simple additive model where each formant contributes independent information to perception based on the discrimination of sounds varying in that formant only, the differences in discrimination found for patterns of parallel and opposing formant movement (i.e., axis orientation in the *APS*) and differences related to the formant patterns of reference points (i.e., location in the *APS*) are not so easily explained. Alternative methods of interpretation must be employed and investigated to understand these differences. For example, a previous discussion has indicated some plausibility for explanations based on the use of spatial frequency dimensions for the expression of changes in spectral patterns. Other models and metrics also exist for representing spectral changes as perceptual distances and

should be considered. In total, the results from Experiment II suggest a number of continuing avenues for future investigation into the realm of discrimination and perception of complex sounds.

Taken as a whole, the experiments presented here may not only both be applied and extended to further understanding of normal auditory processes, but also to a related sphere of research pertaining to the discrimination abilities and deficits of the hearing-impaired. While work in this area has been ongoing for many years, the need for a thorough understanding of auditory and speech processing in the impaired ear have never been greater. The advent of new technology surrounding cochlear implants and digital processing in hearing aids provides the potential for enhancing speech and other auditory signals in real time, a capability only dreamed of not many years ago. Natural extensions of the present experiments investigating differences in target zone boundary perception and discrimination of complex changes in vowel-like sounds between normal and impaired ears are justifiably called for, in that they should provide information beneficial to the further development of speech enhancement methods.

It is the hope of the author that, on the basis of the experiments presented here and the implications for further research stemming from them, the potential for target zone theory and the *APS* conceptual format to serve speech science as a research tool has been demonstrated. Additionally, it is hoped that this work may, in some way, prove to further the advancement of speech science, our knowledge of human perceptual abilities, and our general understanding of the communication process.

Figure 4-1: Locations of *SSB* target zones in APS $x'y'$ coordinates with axes modified to reflect approximately equal DL units.



Bibliography

Ainsworth, W.A. (1972). "Duration as a cue in the recognition of synthetic vowels," J. Acoust. Soc. Am., 51, 648-651.

Ainsworth, W.A. (1975). "Intrinsic and extrinsic factors in vowel judgements," in G. Fant and M. Tatham (eds.) *Auditory Analysis and Perception of Speech.* Academic Press, London, pp. 103-113.

Ainsworth, W.A., and Millar, J.B. (1971). "Methodology of experiments on the perception of synthesized vowels," Language and Speech, 14, 201-212.

Assmann, P., Nearey, T., and Hogan, J. (1982). "Vowel identificaton: Orthographic, perceptual and acoustic aspects," J. Acoust. Soc. Am., 71, 975-989.

Burdick, C.K., and Miller, J.D. (1975). "Speech perception by the chinchilla: discrimination of sustained /a/ and /i/," J. Acoust. Soc. Am., 58, 415-427.

Carlson, R., Granström, B., and Klatt, D. (1979). "Vowel perception: The relative perceptual salience of selected acoustic manipulations," STL-QPSR, 3-4, 73-83.

Clarke, F.R. (1960). "Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests," J. Acoust. Soc. Am., 32, 35-46.

Cohen, J. (1960). "A coefficient of agreement for nominal scales," Ed. and Psych. Measurement, XX, No. 1, 37-46.

Disner, S.F. (1980). "Evaluation of vowel normalization procedures," J. Acoust. Soc. Am., 67, 253-261.

Dunn, H.K. (1961). "Methods of measuring vowel formant bandwidths," J. Acoust. Soc. Am., 33, 1737-1746.

Durlach, N.I., and Braida, L.D. (1969). "Intensity perception: I. Preliminary theory of intensity resolution," J. Acoust. Soc. Am., 46, 372-383.

Fairbanks, G., and Grubb, P. (1961). "A psychophysical investigation of vowel formants," J. Sp. Hear. Res., 4, 203-219.

Fant, G. (1973). *Speech Sounds and Features.* MIT Press, Cambridge, MA., 227 p.

Fant, G. (1972). "Vocal tract wall effects, losses, and resonance bandwidths," STL-QPSR, 2-3, 28-52.

Fant, G. (1967). "Auditory patterns of speech," in W. Wathen-Dunn (ed.) *Models for the Perception of Speech and Visual Form.* MIT Press, Cambridge, MA. pp. 111-125.

Fant, G. (1960). *Acoustic Theory of Speech Production.* Mouton & Co., The Hague, The Netherlands.

Flanagan, J.L. (1955). "A difference limen for vowel formant frequency," J. Acoust. Soc. Am., 27, 613-617.

Fourakis, M.S., and Hawks, J.W. (1990). "On the perceptual vowel space of Modern Greek," J. Acoust. Soc. Am., 87, S159.

Fourakis, M.S., and Miller, J.D. (1987). "Measurement of vowels in isolation and in sonorant context," J. Acoust. Soc. Am., 81, S17.

Fry, D.B., Abramson, A.S., Eimas, P.D., and Liberman, A.M., (1962). "The identification and discrimination of synthetic vowels," Lang. and Speech, 5, 171-189.

Fujimura, O., and Lindqvist, J. (1971). "Sweep-tone measurements of vocal-tract characteristics," J. Acoust. Soc. Am., 49, 541-558.

Fujisaki, H., and Kawashima, T. (1968). "The roles of pitch and the higher formants in the perception of vowels," IEEE Trans. Audio Electroacoust. AU-16, 73-77.

Greenberg, G.Z. (1962). "Cueing signals and frequency uncertainty in auditory detection," Tech. Rep. AF 19(628)-266 U.S.A.F.

Hawks, J.W. and Miller, J.D. (1989). "Perception of synthetic vowels: A comparison of several classification schemes," J. Acoust. Soc. Am., 86, S78.

Holmes, J. (1986). "Normalization in vowel perception," in J. Perkell and D. KLatt (eds.) *Invariance and Variability in Speech Processes.* Lawrence Erlbaum, Hillsdale, NJ, pp. 346-357.

House, A.S., and Stevens, K.N. (1958). "Estimation of formant bandwidths from measurements of transient response of the vocal tract," J. Sp. Hear. Res., 1, 309-315.

Jongman, A., Fourakis, M., and Sereno, J. (1989). "The acoustic vowel space of Modern Greek and German," Lang. and Speech, 32, 221-248.

Kahn, D. (1978). "On the identifiability of isolated vowels," UCLA Working Papers in Phonetics, 41, 26-31.

Kakusho, O., and Kato, K. (1968). "Just discriminable change and matching range of acoustic parameters of vowels," Acustica, 20, 46-54.

Khazatsky, V. (1985). "Computational model of the effects of standards," Unpublished manuscript.

Klatt, D.H. (1977). "A cascade-parallel terminal analog speech synthesizer and a strategy for consonant-vowel synthesis," J. Acoust. Soc. Am., 61, S68.

Klatt, D.H. (1979). "Speech perception: A model of acoustic-phonetic analysis and lexical access," J. Phonetics, 7, 279-312.

Klatt, D.H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., 67, 971-995.

Klatt, D.H. (1982). "Prediction of perceived phonetic distance from critical-band spectra: A first step," Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, 1278-1281.

Klatt, D.H. (1987). "Review of text-to-speech conversion for English," J. Acoust. Soc. Am., 82, 737-757.

Klatt, D.H., and Klatt, L.C. (1990). "Analysis, synthesis and perception of voice quality variaitons among female and male talkers," J. Acoust. Soc.Am., 87, 820-838.

Koenig, W. (1949). "A new frequency scale for acoustic measurements," Bell Labs Record, 27, 299-301.

Kollmeier, B., Gilkey, R.H., and Sieben, U.K. (1988). "Adaptive staircase techniques in psychophysics: A comparison of human data and a mathematical model," J. Acoust. Soc. Am., 83, 1852-1862.

Ladefoged, P. (1982). *A Course in Phonetics.* Harcourt Brace Jovanovich, Inc., New York, NY, 300 p.

Lee, W.A., and Shoup, J.E. (1980). "Specific contribution of the ARPA SUR project," in *Trends in Speech Recognition.* Prentice-Hall, Inc., Englewood Cliffs, NJ.

Lehiste, I., and Peterson, G. (1961). "Transitions, glides, and diphthongs," J. Acoust. Soc. Am., 33, 268-277.

Levitt, H. (1970). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am., 49, 65-69.

Liberman, A.M., Cooper, F.S., Shankweiler, D.S., and Studdert-Kennedy, M. (1967). "Perception of the Speech Code," Psychol. Rev., 74, 431-461.

Liberman, A.M., and Mattingly, I.G. (1985). "The motor theory of speech perception revised," Cognition, 21, 1-36.

Liljencrants J., and Lindblom, B. (1972). "Numerical simulation of vowel quality systems: The role of perceptual contrast," Language, 48, 839-862.

Luce, R.D. (1963). "A threshold theory for simple detection experiments," Psych. Rev., 70, 61-79.

Macmillan, N.A., Braida, L.D., Goldberg, R.F., and Khazatsky, V. (1986). "Central and peripheral processes in the perception of speech and nonspeech sounds," NATO Conference, Utrecht, The Netherlands.

Macmillan, N.A., Goldberg, R.F., and Braida, L.D. (1988). "Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua," J. Acoust. Soc. Am., 84, 1262-1280.

Macmillan, N.A., Kaplan, H.L., and Creelman, C.D., (1977), "The psychophysics of categorical perception," Psych. Rev., 84, 452-471.

McKay, D.M. (1956). "The Epistemological Problems for Automata," in C.E. Shannon and J. McCatthy (eds.) *Automata Studies.* Princeton University Press, Princeton, NJ.

Mermelstein, P. (1978). "Difference limens for formant frequencies of steady-state and consonant-bound vowels," J. Acoust. Soc. Am., 63, 572-580.

Millar, J.B., and Ainsworth, W.A. (1972). "Identification of synthetic isolated vowels and vowels in H-D context," Acustica, 27, 278-282.

Miller, J.D. (1980). "Estimation of formant bandwidths for vowels," Unpublished.

Miller, J.D. (1984). "Auditory processing of the acoustic patterns of speech," Arch. Otolaryngol., 110, 154-159.

Miller, J.D. (1987b). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am., 81, S16.

Miller, J.D. (1987c). "Classifications of vowel productions by means of perceptual target zones: A response to Ladefoged and Studdert-Kennedy," J. Acoust. Soc. Am., 82, S82.

Miller, J.D. (1989). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am., 85, 2114-2121.

Miller, J.D., and Hawks, J.W. (1986). "Spectral envelopes and perceptual target zones for consonants and vowels: Preliminary estimates," J. Acoust. Soc. Am., **79**, S66.

Miller, J.D., and Hawks, J.W. (1989). "Target zones for synthetic vowels," J. Acoust. Soc. Am., **85**, S66.

Miller, R.L. (1953). "Auditory tests with synthetic vowels," J. Acoust. Soc. Am., **18**, 114-121.

Morton, J., and Broadbent, D.E. (1967). "Passive versus active recognition models or is your homunculus really necessary?" in W. Wathen-Dunn (ed.) *Models for the Perception of Speech and Visual Form.* MIT Press, Cambridge, MA. pp. 103-110.

Neary, T.M. (1977). "Phonetic feature systems for vowels," Ph.D. Thesis, Univ. of Connecticut. Reproduced by Indiana University Linguistics Club, 1978.

Neary, T.M. (1989). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am., **85**, 2088-2113.

Nord, L., and Sventelius, E. (1979). "Analysis and prediction of difference limen data for formant frequencies," STL-QPSR, **3-4**, 60-72.

Peterson, G.E., and Barney, H.L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am., **24**, 175-184.

Peterson, G.E., and Lehiste, I. (1960). "Duration of syllable nuclei in English," J. Acoust. Soc. Am., **32**, 693-703.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (1986). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech, J. Sp. Hear. Res., **29**, 434-446.

Pike, K. (1947). "On the phonemic status of English diphthongs," Language, **23**, 151-159.

Pisoni, D.B. (1971). "On the Nature of categorical perception of speech sounds." Ph.D. Thesis. University of Michigan.

Pisoni, D.B. (1973). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Perception and Psychophysics, , 13, 253-260.

Plomp, R. (1970). "Timbre as a multidimensional attribute of complex tones, " in R. Plomp and G.F. Smoorenburg (Eds.) *Frequency Analysis and Periodicity Detection in Hearing.* Sijthoff, Leiden, The Netherlands, pp. 397-414.

Pollack, I., and Decker, L.R. (1958). "Confidence ratings, message reception and the receiver operating characteristic," J. Acoust. Soc. Am., 30, 286-292.

Pols, L.C.W., Van Der Kamp L. J. Th., and Plomp, R. (1969). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am., 46, 458-467.

Pols, L.C.W., Tromp, H.R.C., and Plomp, R. (1973). "Frequency analysis of Dutch vowels from 50 male speakers," J. Acoust. Soc. Am., 53, 1093-1101.

Potter, R.K., and Peterson, G.E. (1948). "The representation of vowels and their movements," J. Acoust. Soc. Am., 20, 528-535.

Repp, B.H., Healy, A.F., and Crowder, R.G. (1979). "Categories and context in the perception of isolated steady-state vowels," J. Exp. Psychol.; Human Perception and Performance, 5, 129-145.

Robinson, D.E., and Watson, C.S. (1972). "Psychophysical methods in modern psychoacoustics," In J.V. Tobias (Ed.) *Foundations of Modern Auditory Theory.* Academic Press, New York, NY, pp. 101-131.

Sachs, R.M. (1969). "Vowel identification and discrimination in isolation vs. word context," Quarterly Progress Report No. 93, Research Laboratory of Electronics, M.I.T., 220-229.

Scholes, R.J. (1967). "Categorial responses to synthetic vocalic stimuli by speakers of various languages," Lang. and Speech, 10, 252-282.

Stevens, K.N. (1968). "On the relations between speech movements and speech perception," Zeitschr. f. Phon., 21, 102.

Stevens, K.N., and Halle, M. (1967). "Remarks on analysis by synthesis and distinctive features," in W. Wathen-Dunn (ed.) *Models for the Perception of Speech and Visual Form.* MIT Press, Cambridge, MA. pp. 88-102.

Stevens, K.N., Liberman, A.M., Studdert-Kennedy, M., and Öhman, S.E.G. (1969). "Crosslanguage study of vowel perception," Lang. and Speech, **12**, 1-23.

Strange, W. (1989). "Evolving theories of vowel perception," J. Acoust. Soc. Am., **85**, 2081-2087.

Strange, W., Edman, T.R., and Jenkins, J.J. (1979). "Acoustic and phonological factors in vowel identification," J. Exp. Psych.: Human Perception and Performance, **5**, 643-656.

Strange, W., Verbrugge, R., Shankweiler, D., and Edman, T. (1976). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am., **60**, 213-224.

Syrdal, A., and Gopal, H. (1986). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am., **79**, 1086-1100.

Tatsuoka, M. (1970). *Selected Topics in Advanced Statistics: An Elementary Approach Pt. 6: Discrimination Analysis* (Institute for Personality and Ability Testing, Champaign, IL).

Taylor, M.M., and Creelman, C.D. (1967). "PEST: Efficient estimates on probability functions," J. Acoust. Soc. Am., **41**, 782-787.

Traunmüller, H. (1981). "Perceptual dimensions of openness in vowels," J. Acoust. Soc. Am., **69**, 1465-1475.

Zwicker, E., and Terhardt, E. (1980). "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am., **68**, 1523-1525.

Appendix A

Formant location in the *APS*

This section will demonstrate and discuss how certain formant patterns manifest themselves in x', y', z' space and will be limited to cases where the sensory reference (SR) remains fixed. Recall from section 1.2.1 that the auditory-perceptual space (APS) is defined in three dimensions, x, y, and z, where

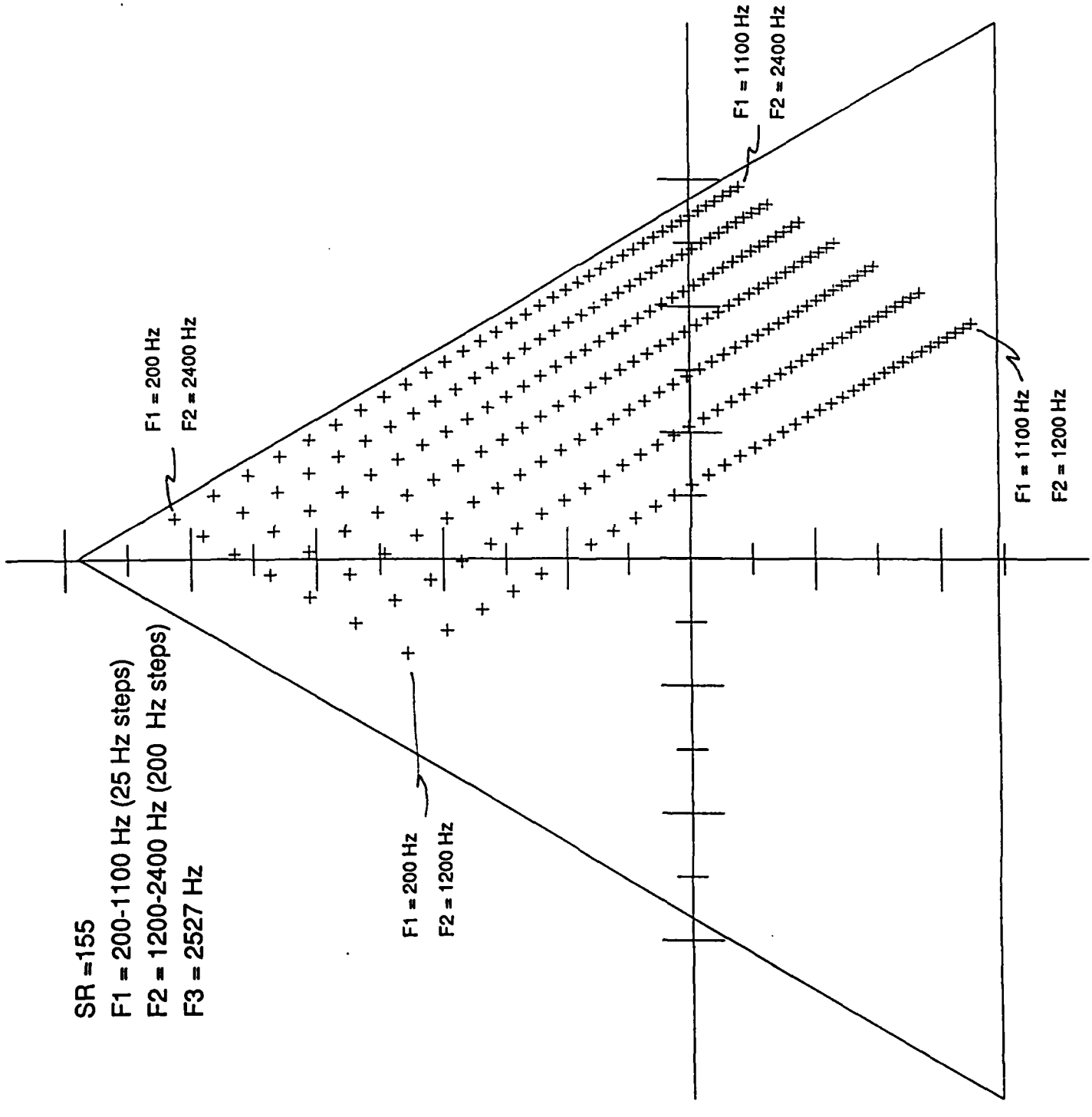
$$\begin{aligned} x &= log(SF3/SF2), \\ y &= log(SF1/SR), \quad and \\ z &= log(SF2/SF1). \end{aligned} \tag{A.1}$$

These coordinates are often transformed for visual simplification by rotation the APS axes, yielding a new set of coordinates, x', y', z', where,

$$\begin{aligned} x' &= .70711(y - x), \\ y' &= .8162(z) - .4081(x + y), and \\ z' &= .5772(x + y + z). \end{aligned} \tag{A.2}$$

Figure A-1 illustrates the location of points generated in $x'y'$-space when $F2$ and $F3$ are fixed and only $F1$ is allowed to vary. The figure shows seven examples of this configuration with different values of $F2$. Such configurations might be approximated in natural speech as formant movements from /AE/ to /IY/ or /AA/ to /UW/. As $F1$ increases, points associated with a given $F2$ move down and to the right creating a line of points which lies at a 60° angle relative to the x' axis. These lines represent constant values of $F2$. All points lie in

Figure A-1: Location of seven continua generated in $x'y'$ space with fixed values of $F2$ and $F3$ with $F1$ allowed to vary.



a single z' plane ($z' = 0.70$) and, since $F1$ is incremented in linear steps, are logarithmically spaced. Figure A-2 demonstrates somewhat the opposite angular effect. Here $F1$ and $F3$ are fixed and only $F2$ allowed to vary. Once again seven examples are illustrated with different values of $F1$. Similar configurations to these might be approximated in natural speech as formant movements from /UW/ to /IY/ or /IH/, or /AO/ to /AE/ or /EH/. All points still lie in a single z' plane and the points associated with a given $F1$ form a line now at a 120° angle to the x' axis.

If $F1$ and $F2$ are held constant and $F3$ is allowed to vary, the patterns seen in Figure A-3 emerge. These patterns might be similar to paths seen when the back vowels /UW/ or /AO/ are roticized, as when they precede /R/ in natural speech. Two groups of data points are presented with one group representing a fixed value for $F1$ with several values of $F2$ and another group representing a fixed value of $F2$ with several values of $F1$. Irrespective of the groupings, a line of points lying at a 150° angle relative to the x' axis is generated for each pair of $F1$ and $F2$ values. Figure A-4 provides a "side view" of the same data points in $y'z'$-coordinate space. The lines of points move "forward" along z' (left to right in this view) at a 33° angle to the z' axis as $F3$ increases. Thus the "lines" of points for pairs of $F1$ and $F2$ seen in Figure A-3 are somewhat deceiving in that the movement is occuring along the z' axis, and therefore, for any given value of z' only a single point would be present for each formant pair.

If $F1$ and $F2$ maintain a constant ratio with each other (with $F3$ fixed), approximating formant movements from /UW/ to /AE/ in natural speech, points of like ratios form lines falling parallel to the x' axis. This is illustrated in Figure A-5 which shows eight horizontal lines of points where $log(F1/F2) = 0.10$ to 0.38 in 0.04 steps. However, these log ratios may also be expressed as simple formant ratios where $F2/F1 = 1.26$ to 2.4. Once again all points lie stationary in one z' plane.

Figure A-6 demonstrates the effect of formants moving toward or away from each other. In this figure, the lowest point represents a merged $F1$ and $F2$, similar to /AA/, with data points moving upward along the y' axis as $F1$ and $F2$ move away from each other, to an /IH/ or /IY/ configuration. Again, with $F3$ fixed, there is no movement along the z' axis. For lines of data points to lie parallel to the y' axis and have no movement along z', the

Figure A-2: Location of seven continua generated in $x'y'$ space with fixed values of $F1$ and $F3$ with $F2$ allowed to vary.

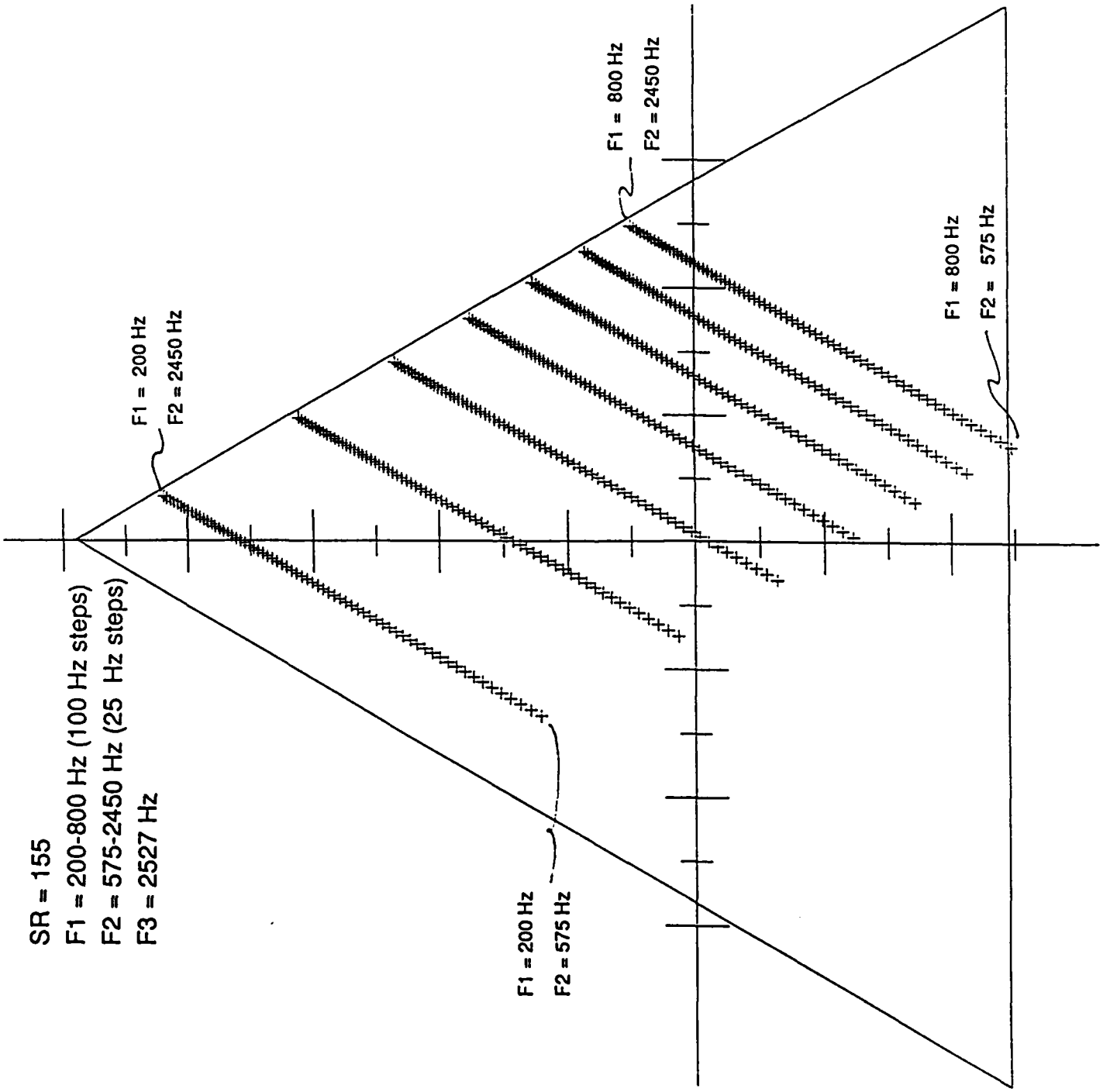


Figure A-3: Location of seven continua generated in $x'y'$ space with fixed values of $F1$ and $F2$ with $F3$ allowed to vary. Crosses indicate continua with a fixed $F1$ and $F2$ changing with each continuum. Squares indicate continua with a fixed $F2$ and $F1$ changing with each continuum.

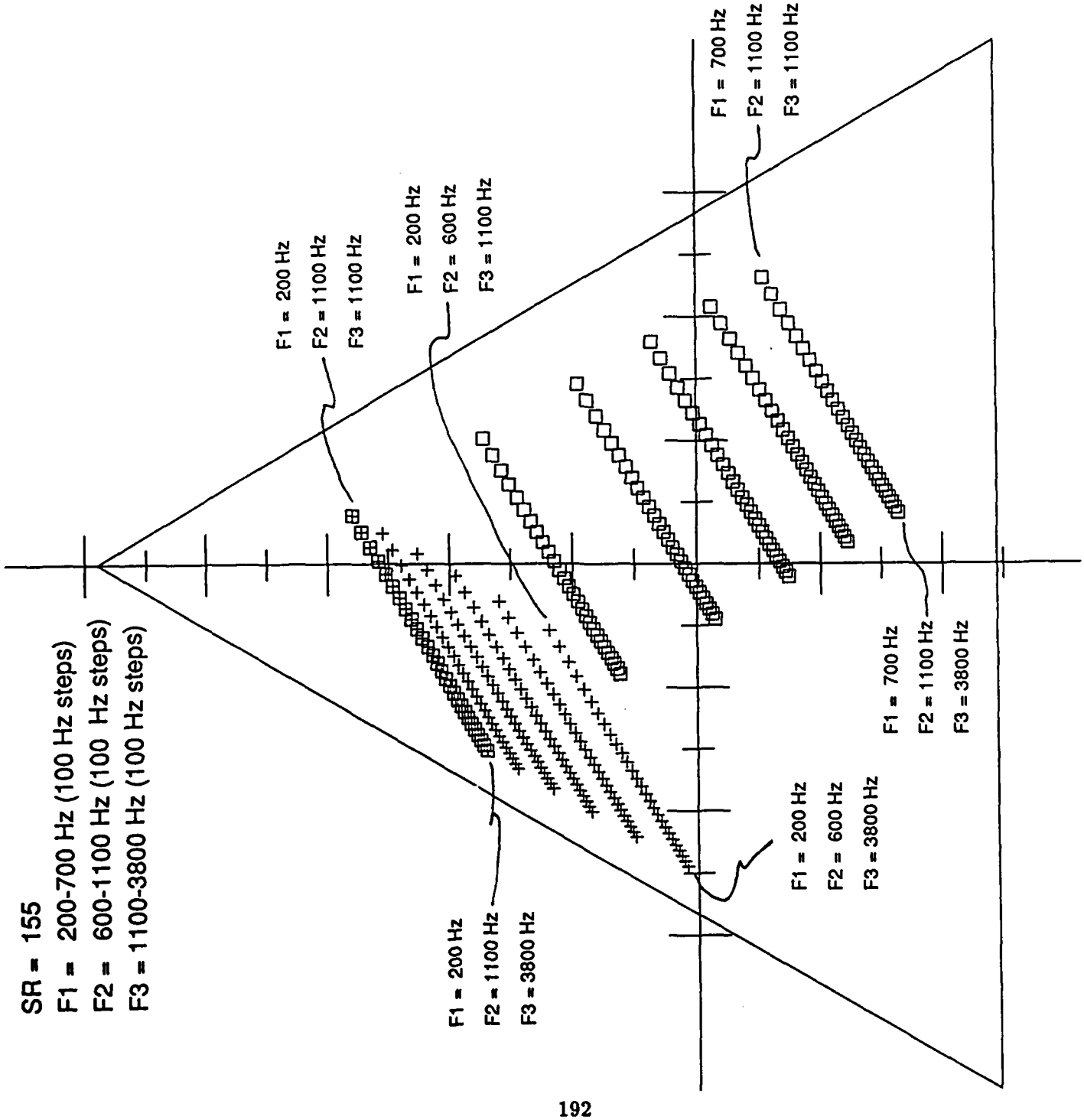


Figure A-4: "Side" view in $y'z'$ space of continua from Figure A-3.

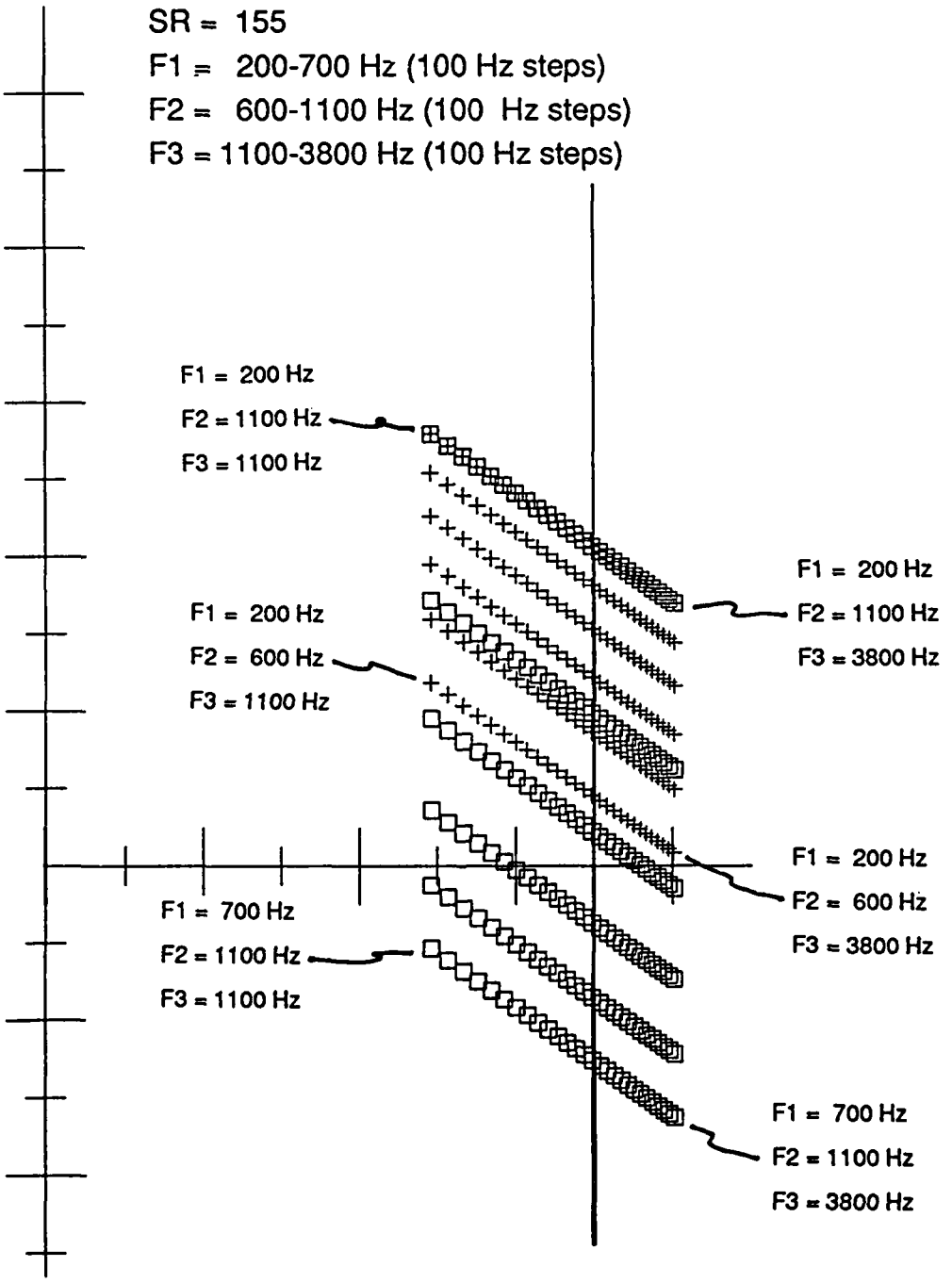


Figure A-5: Location of eight continua generated in $x'y'$ space with a fixed $F3$ and $F1$ and $F2$ maintained in constant ratios.

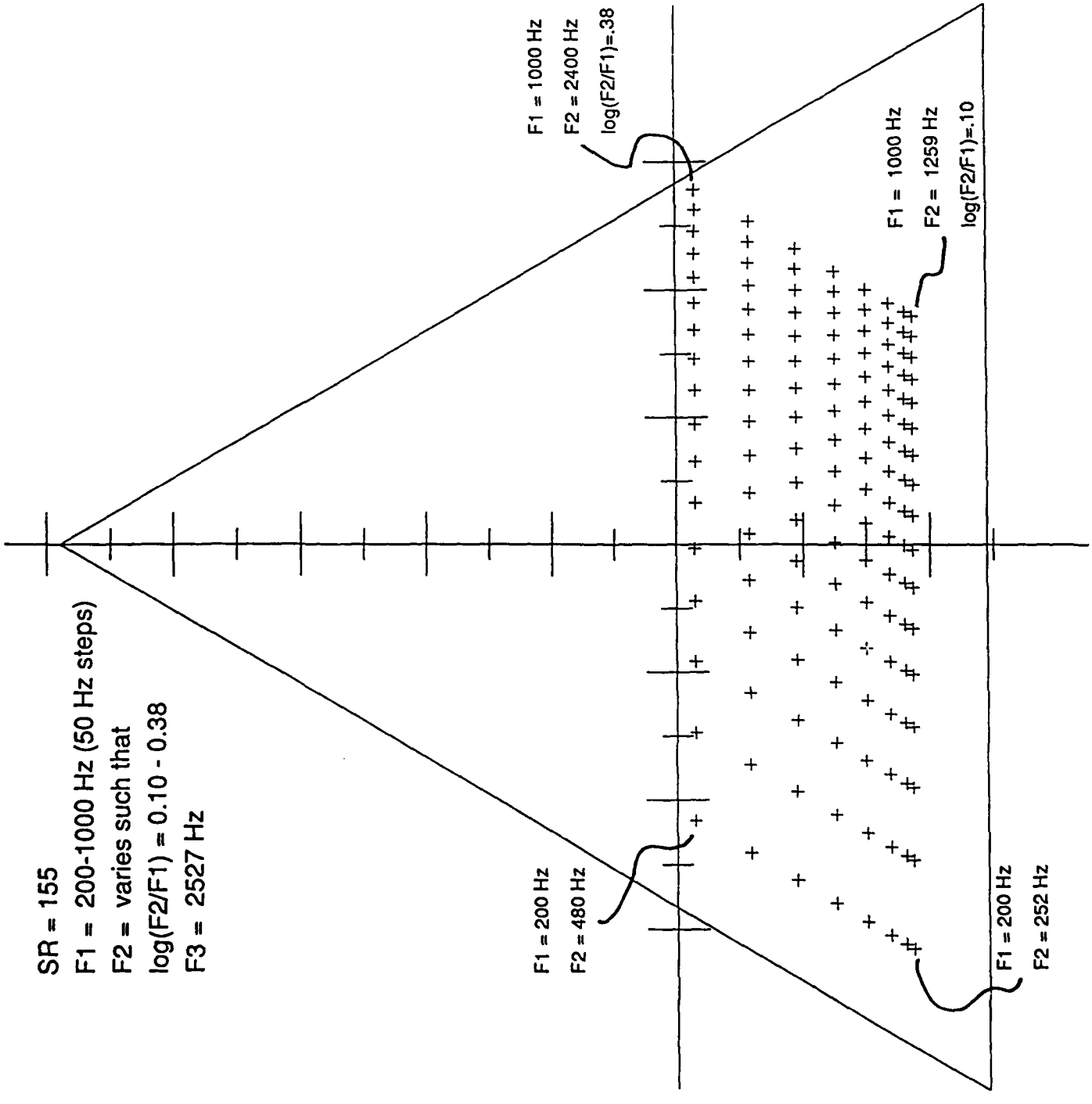
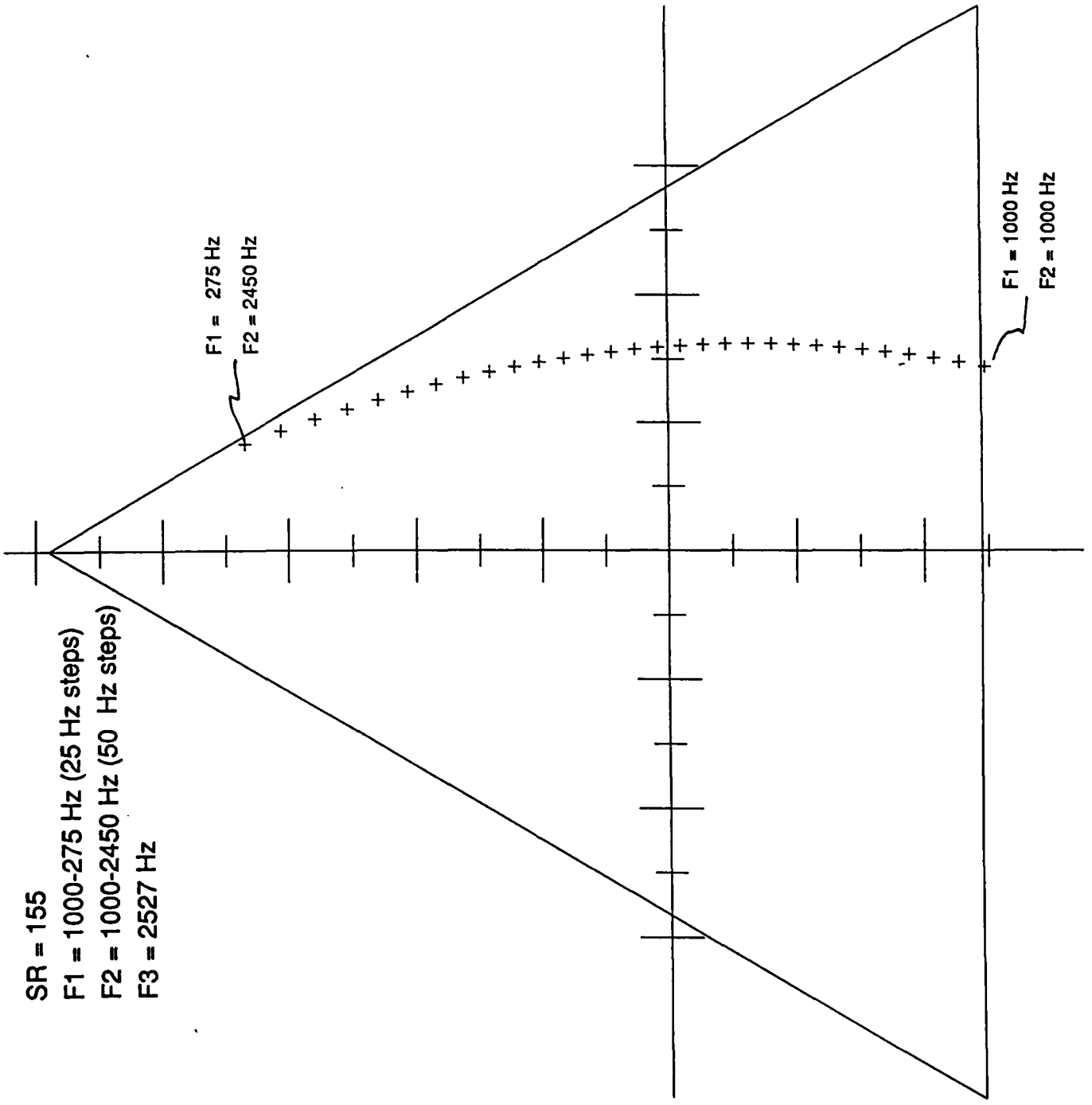


Figure A-6: Location of continuum generated in $x'y'$ space with $F'3$ fixed and increasingly greater separation in $F1$ and $F2$.



values of $F1$ and $F2$ must be manipulated such that

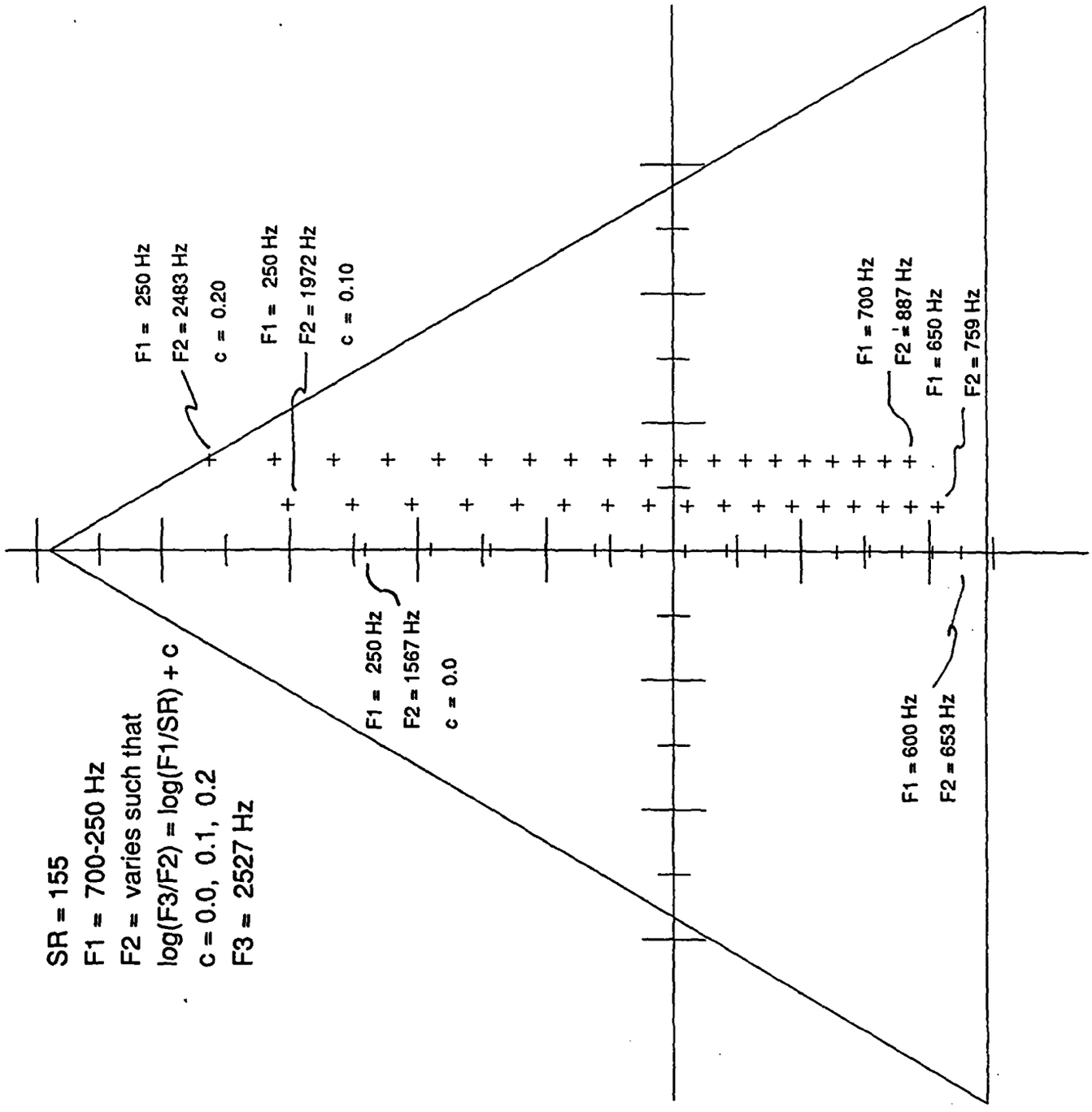
$$log(F3/F2) = log(F1/SR) + c, \tag{A.3}$$

where $F3$, SR, and c are constants. This is illustrated in Figure A-7 with each of the three vertical rows of points reflecting a different value of c. Although this parallel movement along the y' axis appears complicated, the more general aspect of movement along this axis is that $F1$ and $F2$ are moving in opposite directions.

In summary, points generated with constant values of $F1$ fall along 120° lines relative to the x' axis with $F1$ increasing top-left to bottom-right in the $x'y'$ view. Points generated with constant values of $F2$ fall along 60° lines relative to the x' axis with $F2$ increasing bottom-left to top-right in the $x'y'$ view. Points genreated with constant values of $F3$ fall along lines which are perpendicular planes to the z' axis with $F3$ increasing "back to front" in the $x' - y'$ view. When both $F1$ and $F2$ are held constant, representative points fall along lines appearing to fall at a 150° angle to the x' axis in the $x'y'$ view, but are actually angular lines along the z' axis. When $F1$ and $F2$ both increase or decrease together, representative points fall along lines more parallel to the x' axis, while points representing values of $F1$ and $F2$ moving toward or away from each other fall along lines more parallel to y'.

The three-dimensional metric utilized by the *APS* provides a valuable and innovative method of visualizing changes in complex sounds. Although formant change in natural speech does not behave as precisely or as simply as those shown in the preceding figures, it is hoped that this section has provided a better understanding of how the locations and changes in speech sounds may be interpreted with such a metric.

Figure A-7: Location of three continua generated in $x'y'$ space parallel to the y' axis. SR, $F3$, and a constant c are fixed.



Appendix B

Synthesis Parameter Specifications

All parameters specified for the synthesis of stimuli are shown in Table B.1 as they appear in the synthesizer program. The columns under "SYM" indicate show the abbreviations for the parameters with the "v/C" columns indicating whether that parameter is of a variable or constant specification. The columns under "VAL" indicate values used for token synthesis with the exception of parameters F1, F2, F3, b1, b2, b3, g0 (overall gain) which could vary with each token, and parameters f0 and av, which are identical for each token, vary over time. The variations over time for these two parameters are shown in Table B.2. For those unfamiliar with the Klatt synthesis program, definitions for most variables can be found in Klatt, 1980; 1987 and Klatt and Klatt, 1990.

Bandwidths in Hertz by formant frequency based on data from Miller(1980) are shown in Table B.3. These values were used for the specification of bandwidths for $F1$, $F2$, and $F3$ for all synthetic tokens generated for both experiments. Formant frequencies falling between values specified in the table were assigned the lower bandwidth value. See Section 2.2.1 for additional details on formant bandwidth calculation.

Table B.1: Synthesis parameter specifications.

SYM	v/C	MIN	VAL	MAX	SYM	v/C	MIN	VAL	MAX
sr	C	5000	10000	20000	nf	C	1	4	8
du	C	30	400	5000	ss	C	1	2	2
ui	C	1	5	20	rs	C	1	1	99999
f0	v	0	1000	5000	av	v	0	60	80
F1	v	180	270	1300	b1	v	30	49	1000
F2	v	550	2290	3000	b2	v	40	105	1000
F3	v	1139	1139	4800	b3	v	60	152	1000
F4	v	2400	4000	4990	b4	v	100	500	1000
F5	v	3000	4900	4990	b5	v	100	1000	1500
f6	v	3000	4990	4990	b6	v	100	500	4000
fz	v	180	280	800	bz	v	40	90	1000
fp	v	180	280	500	bp	v	40	90	1000
ah	v	0	0	80	oq	v	10	50	80
at	v	0	0	80	tl	v	0	0	34
af	v	0	0	80	sk	v	0	0	100
a1	v	0	0	80	p1	v	30	80	1000
a2	v	0	0	80	p2	v	40	200	1000
a3	v	0	0	80	p3	v	60	350	1000
a4	v	0	0	80	p4	v	100	500	1000
a5	v	0	0	80	p5	v	100	600	1500
a6	v	0	0	80	p6	v	100	800	4000
an	v	0	0	80	ab	v	0	0	80
ap	v	0	0	80	os	C	0	0	20
g0	v	0	68	80	dF	v	0	0	100
db	v	0	0	400					

Table B.2: Time-varying synthesis parameter specifications for F0 (x10) and amplitude.

Time	F0	AV	Time	F0	AV	Time	F0	AV
0	1140	1	135	1320	55	270	1205	55
5	1156	6	140	1320	55	275	1196	55
10	1172	12	145	1320	55	280	1188	55
15	1189	17	150	1320	55	285	1180	55
20	1205	23	155	1320	55	290	1172	55
25	1221	28	160	1320	55	295	1164	55
30	1238	33	165	1320	55	300	1155	55
35	1254	39	170	1320	55	305	1147	55
40	1270	44	175	1320	55	310	1139	55
45	1287	50	180	1320	55	315	1131	55
50	1303	55	185	1320	55	320	1123	55
55	1320	55	190	1320	55	325	1114	55
60	1320	55	195	1320	55	330	1106	55
65	1320	55	200	1320	55	335	1098	55
70	1320	55	205	1311	55	340	1090	55
75	1320	55	210	1303	55	345	1082	55
80	1320	55	215	1295	55	350	1073	54
85	1320	55	220	1287	55	355	1065	53
90	1320	55	225	1278	55	360	1057	52
95	1320	55	230	1270	55	365	1049	50
100	1320	55	235	1262	55	370	1041	48
105	1320	55	240	1254	55	375	1032	44
110	1320	55	245	1246	55	380	1024	39
115	1320	55	250	1237	55	385	1016	31
120	1320	55	255	1229	55	390	1008	19
125	1320	55	260	1221	55	395	1000	1
130	1320	55	265	1213	55			
135	1320	55	270	1205	55			

Table B.3: Formant bandwidths (BW) by formant frequency (Frmt) in Hertz utilized for all synthetic tokens in all experiments.

Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW
133	71	1233	63	2238	103	2887	143	3385	183	3799	223
134	70	1267	64	2257	104	2901	144	3396	184	3808	224
136	69	1300	65	2276	105	2915	145	3407	185	3818	225
138	68	1333	66	2295	106	2929	146	3418	186	3827	226
140	67	1365	67	2313	107	2942	147	3429	187	3837	227
143	66	1396	68	2332	108	2956	148	3440	188	3846	228
145	65	1427	69	2350	109	2969	149	3451	189	3856	229
148	64	1457	70	2368	110	2983	150	3462	190	3865	230
151	63	1487	71	2386	111	2996	151	3473	191	3874	231
154	62	1516	72	2404	112	3009	152	3484	192	3884	232
158	61	1545	73	2421	113	3022	153	3495	193	3893	233
161	60	1573	74	2439	114	3035	154	3505	194	3902	234
166	59	1600	75	2456	115	3048	155	3516	195	3911	235
170	58	1628	76	2473	116	3061	156	3527	196	3921	236
175	57	1654	77	2490	117	3074	157	3537	197	3930	237
181	56	1681	78	2507	118	3087	158	3548	198	3939	238
188	55	1707	79	2523	119	3100	159	3558	199	3948	239
195	54	1732	80	2540	120	3112	160	3569	200	3957	240
203	53	1757	81	2556	121	3125	161	3579	201	3966	241
213	52	1782	82	2573	122	3137	162	3589	202	3975	242
225	51	1807	83	2589	123	3150	163	3600	203	3984	243
240	50	1831	84	2605	124	3162	164	3610	204	3993	244
259	49	1855	85	2621	125	3174	165	3620	205	4002	245
286	48	1878	86	2636	126	3186	166	3631	206	4011	246
332	47	1902	87	2652	127	3199	167	3641	207	4020	247
512	48	1925	88	2667	128	3211	168	3651	208	4029	248
600	49	1947	89	2683	129	3223	169	3661	209	4038	249
668	50	1970	90	2698	130	3235	170	3671	210	4047	250
727	51	1992	91	2713	131	3247	171	3681	211	4056	251
780	52	2013	92	2728	132	3258	172	3691	212	4064	252
830	53	2035	93	2743	133	3270	173	3701	213	4073	253
878	54	2056	94	2758	134	3282	174	3711	214	4082	254
923	55	2077	95	2773	135	3294	175	3721	215	4091	255
966	56	2098	96	2788	136	3305	176	3731	216	4099	256
1008	57	2119	97	2802	137	3317	177	3740	217	4108	257
1048	58	2139	98	2817	138	3328	178	3750	218	4116	258
1087	59	2159	99	2831	139	3340	179	3760	219	4125	259
1125	60	2179	100	2845	140	3351	180	3770	220	4134	260
1162	61	2199	101	2859	141	3362	181	3779	221	4142	261
1198	62	2219	102	2873	142	3374	182	3789	222	4151	262

Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW
4159	263	4317	282	4468	301	4612	320	4751	339	4885	358
4168	264	4325	283	4476	302	4620	321	4758	340	4892	359
4176	265	4333	284	4483	303	4627	322	4766	341	4899	360
4185	266	4341	285	4491	304	4635	323	4773	342	4905	361
4193	267	4350	286	4499	305	4642	324	4780	343	4912	362
4202	268	4358	287	4507	306	4650	325	4787	344	4919	363
4210	269	4366	288	4514	307	4657	326	4794	345	4926	364
4218	270	4374	289	4522	308	4664	327	4801	346	4933	365
4227	271	4382	290	4530	309	4672	328	4808	347	4940	366
4235	272	4389	291	4537	310	4679	329	4815	348	4946	367
4243	273	4397	292	4545	311	4686	330	4822	349	4953	368
4252	274	4405	293	4552	312	4694	331	4829	350	4960	369
4260	275	4413	294	4560	313	4701	332	4836	351	4967	370
4268	276	4421	295	4568	314	4708	333	4843	352	4973	371
4276	277	4429	296	4575	315	4715	334	4850	353	4980	372
4285	278	4437	297	4583	316	4723	335	4857	354	4987	373
4293	279	4445	298	4590	317	4730	336	4864	355	4993	374
4301	280	4452	299	4598	318	4737	337	4871	356	5000	375
4309	281	4460	300	4605	319	4744	338	4878	357		
4317	282	4468	301	4612	320	4751	339	4885	358		

Appendix C

Spectral Envelopes for Experiment II reference tokens

Table C.1 lists the values of $F1$, $F2$, and $F3$ for the reference tokens used in Experiment II. The figures that follow represent the spectral envelopes of the 17 reference point tokens utilized in Experiment II. The spectral envelopes for the center reference points can be considered as representive of tokens used in Experiment I as well. These envelopes are the results of 256-point FFTs centered at the 120 msec point of each token and based on the LPC coefficients generated from analyses with the ILS software package. The LPC analyses computed 12 coefficients based on a 25.6 msec hamming window shifted in 1 msec steps along the pre-emphasized signal. The vertical lines in the spectral envelopes represent points of measurement and not harmonic content. The figures are labeled in a fashion similar to that used in Figures 3-2 and 3-5.

Table C.1: Formant ($F1$, $F2$, $F3$) values for the 17 reference points used in Experiment II.

Reference	$F1$	$F2$	$F3$
IY	247	2275	3086
IH	411	1830	2528
EH	583	1785	2528
AE	828	1740	2528
AA	921	1329	2528
AH	627	1199	2528
AO	643	844	2528
UH	388	1188	2528
UW	227	764	2528
ER	401	1213	1390
IYIH	316	2089	2528
IHEH	475	1925	2528
EHAE	652	2072	2528
AEAH	739	1553	2528
AHAA	663	1133	2528
AHUH	459	1143	2528
UHUW	329	1046	2528

Figure C-1: Spectral envelope derived from FFT of [IY] reference token.

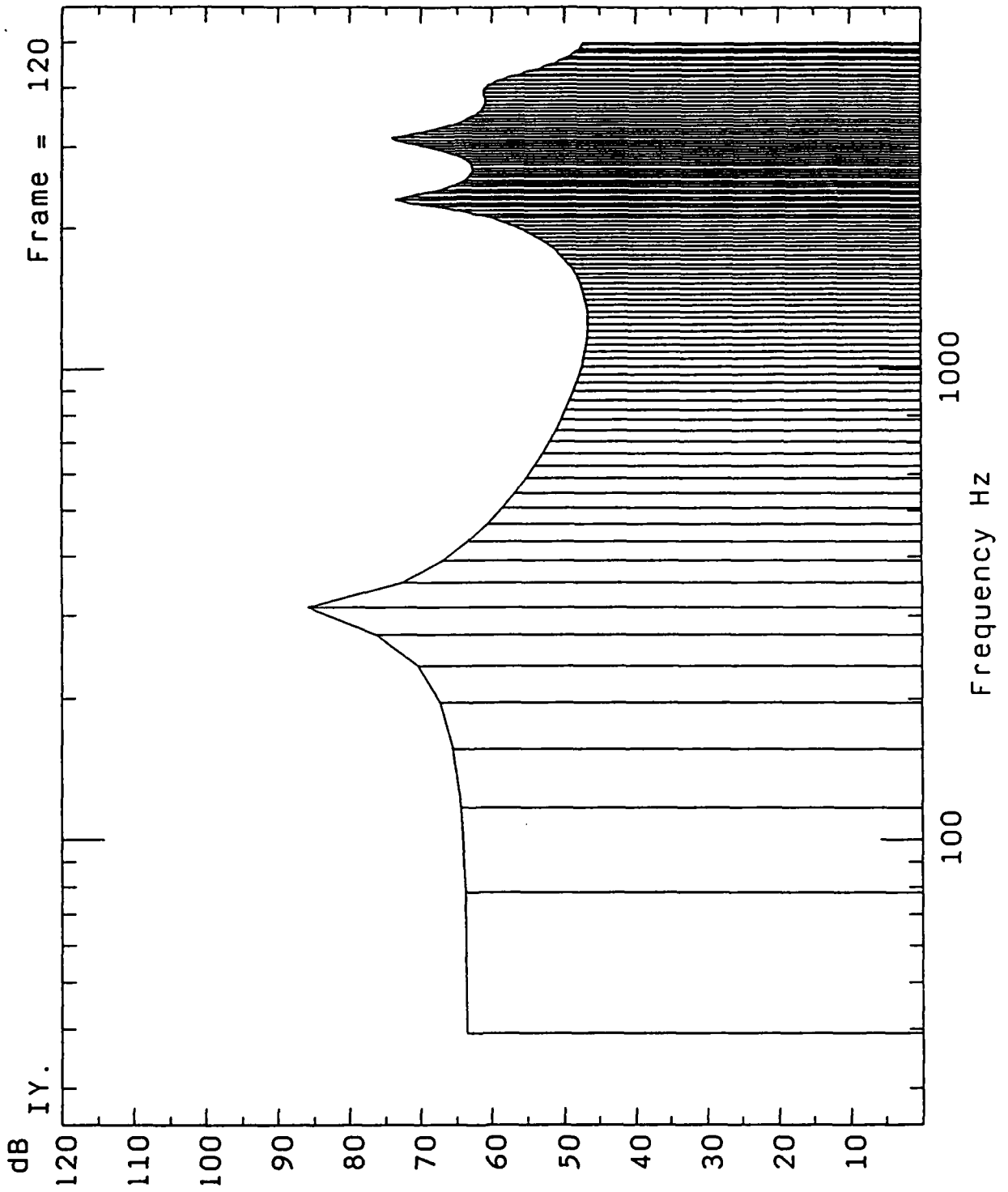


Figure C-2: Spectral envelope derived from FFT of [IH] reference token.

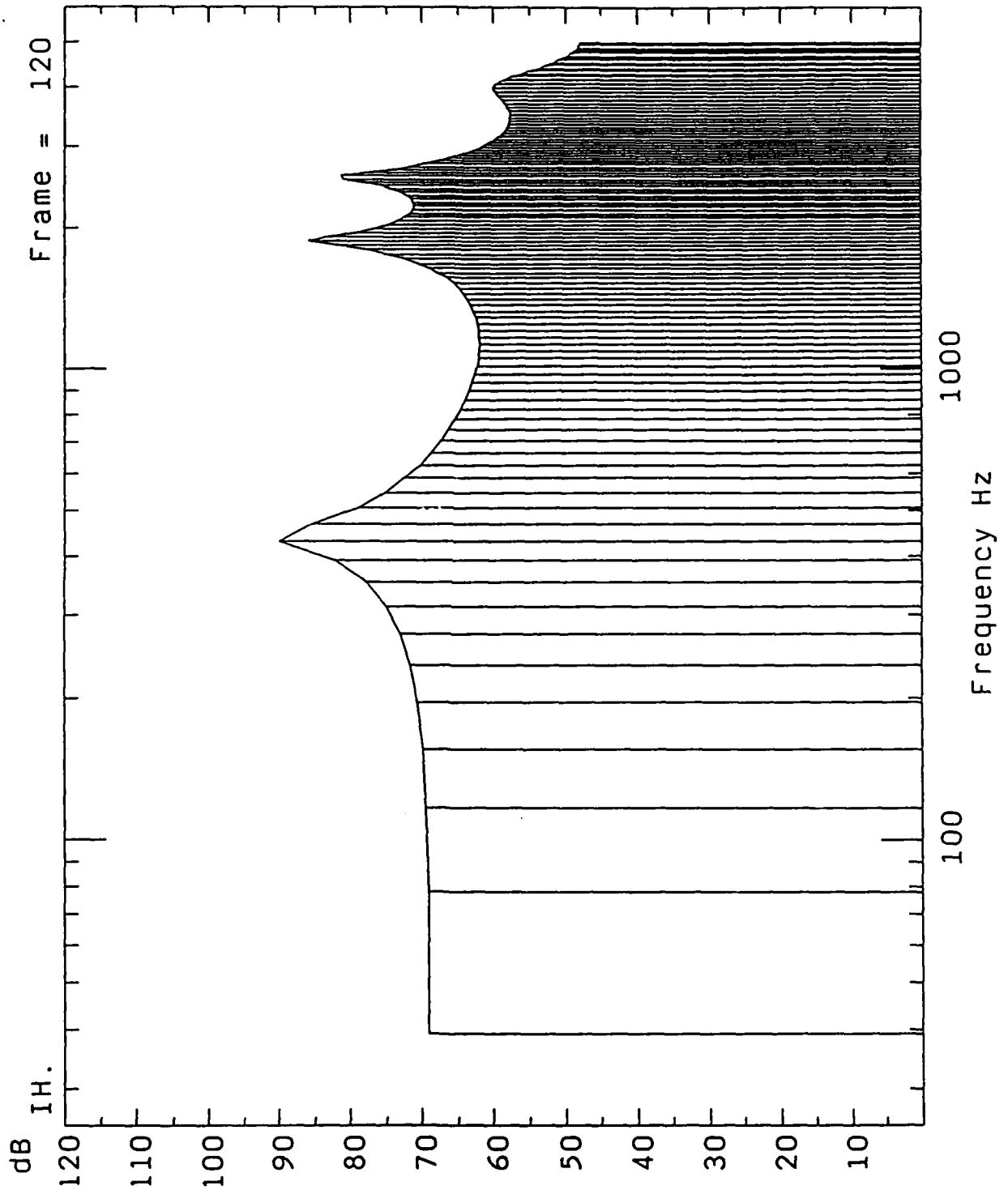


Figure C-3: Spectral envelope derived from FFT of [EH] reference token.

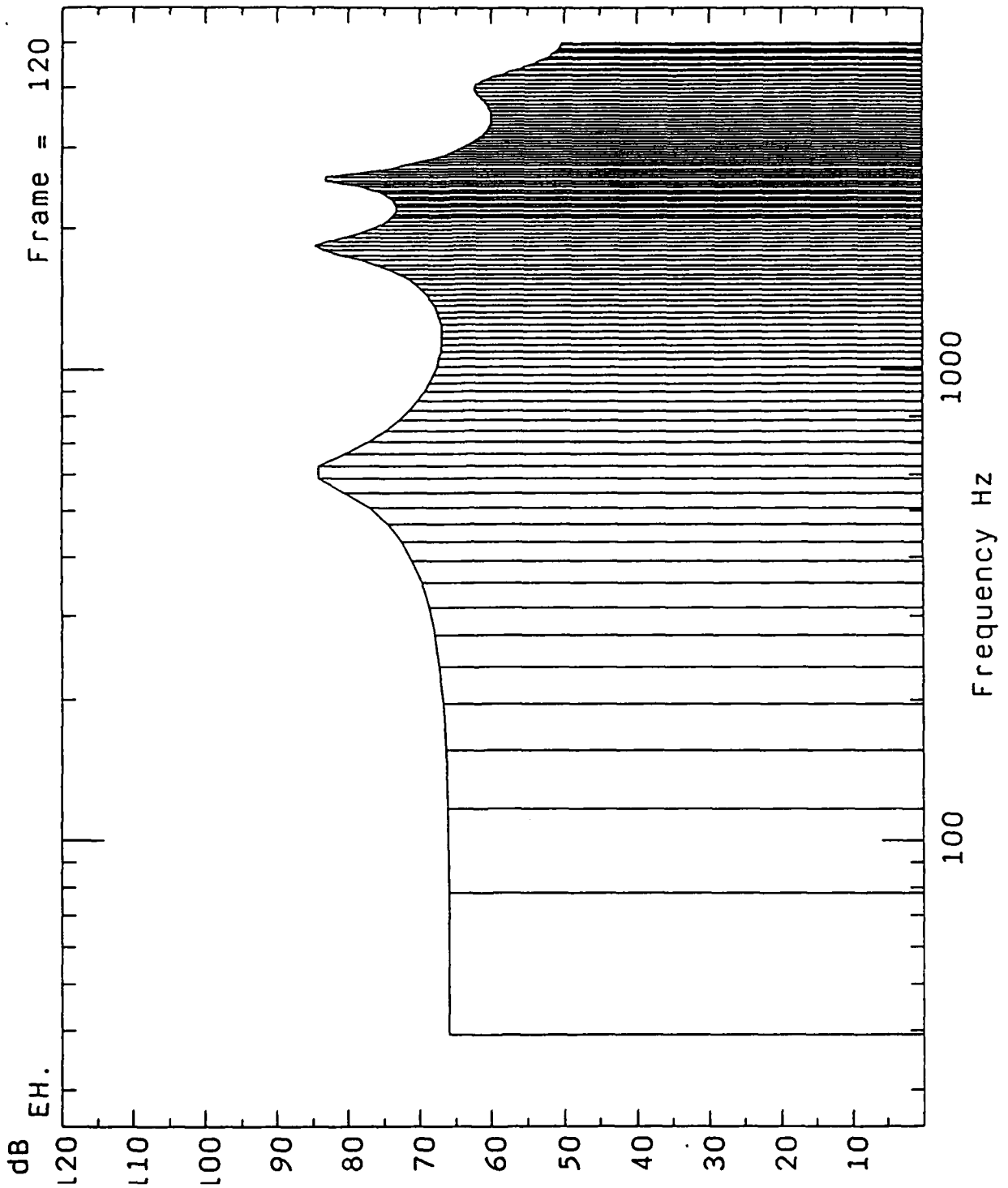


Figure C-4: Spectral envelope derived from FFT of [AE] reference token.

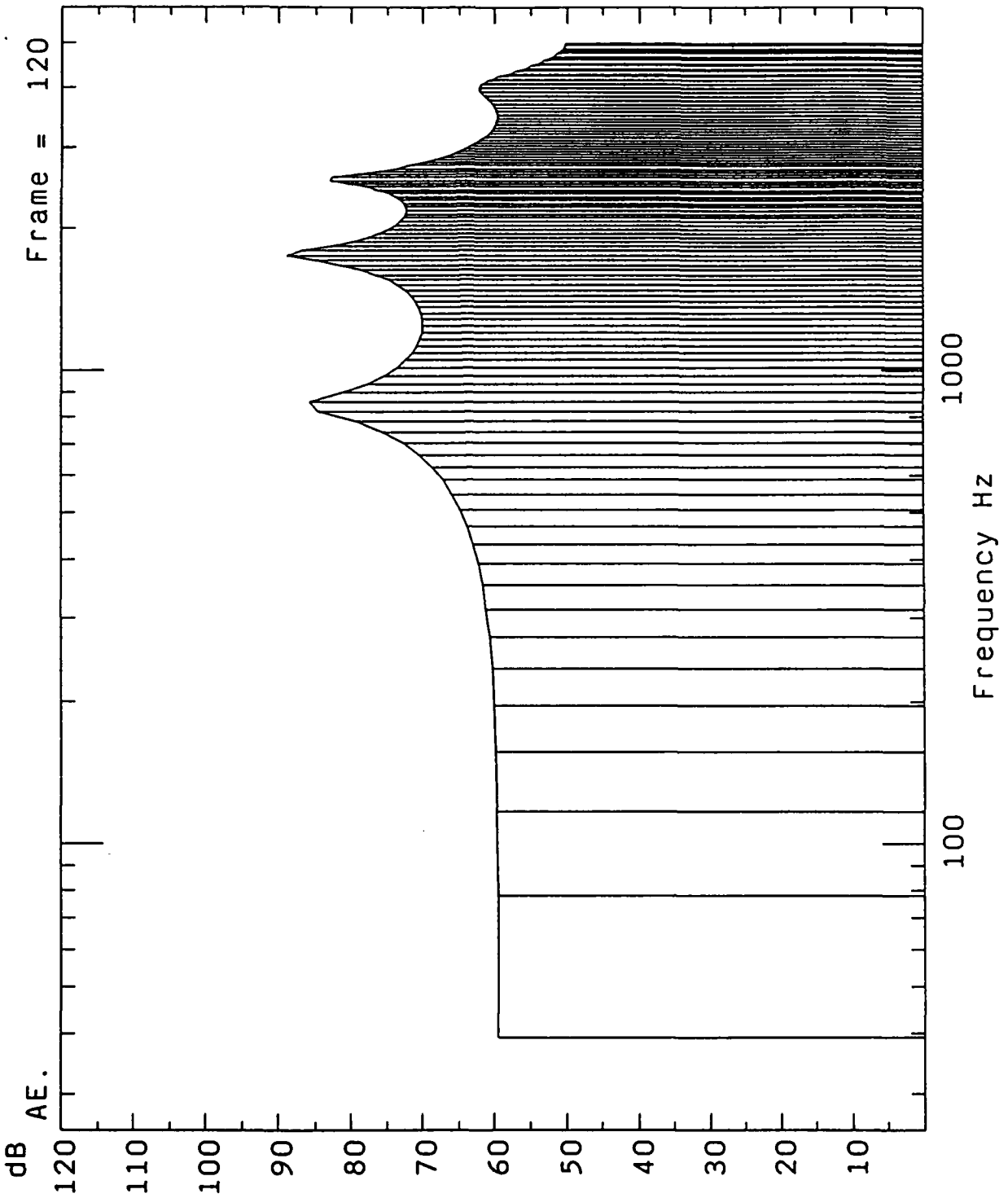


Figure C-5: Spectral envelope derived from FFT of [AA] reference token.

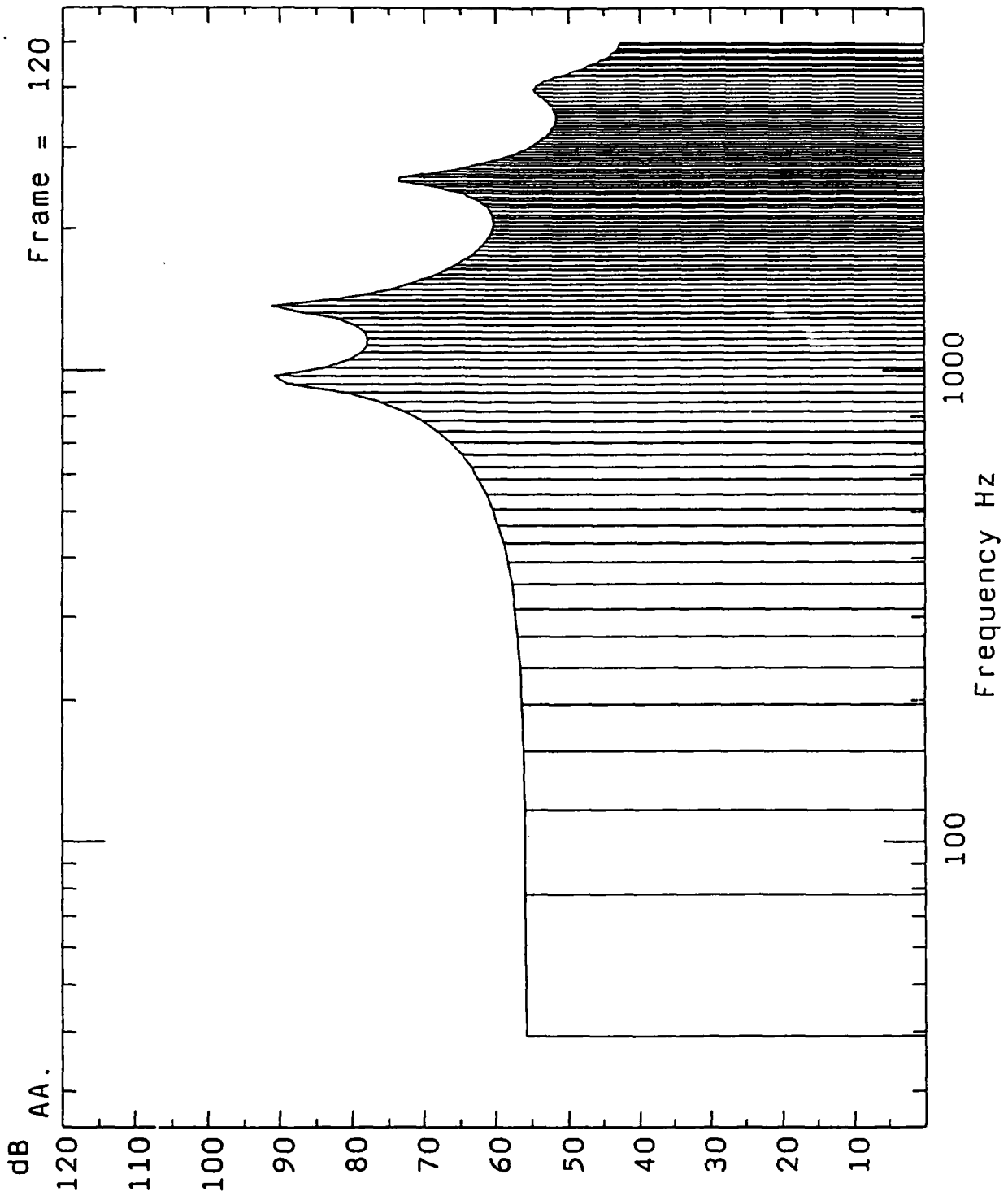


Figure C-6: Spectral envelope derived from FFT of [AO] reference token.

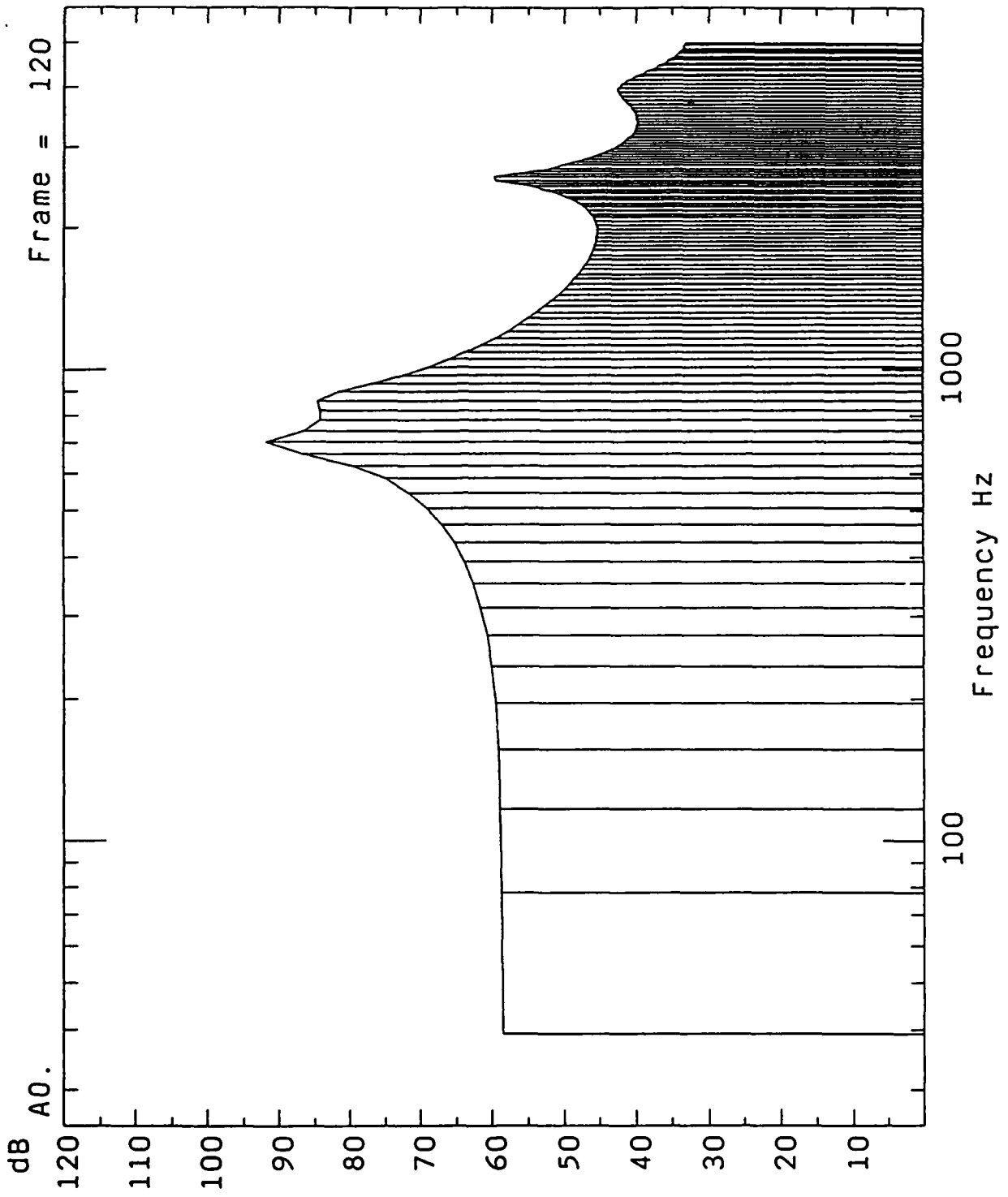


Figure C-7: Spectral envelope derived from FFT of [AH] reference token.

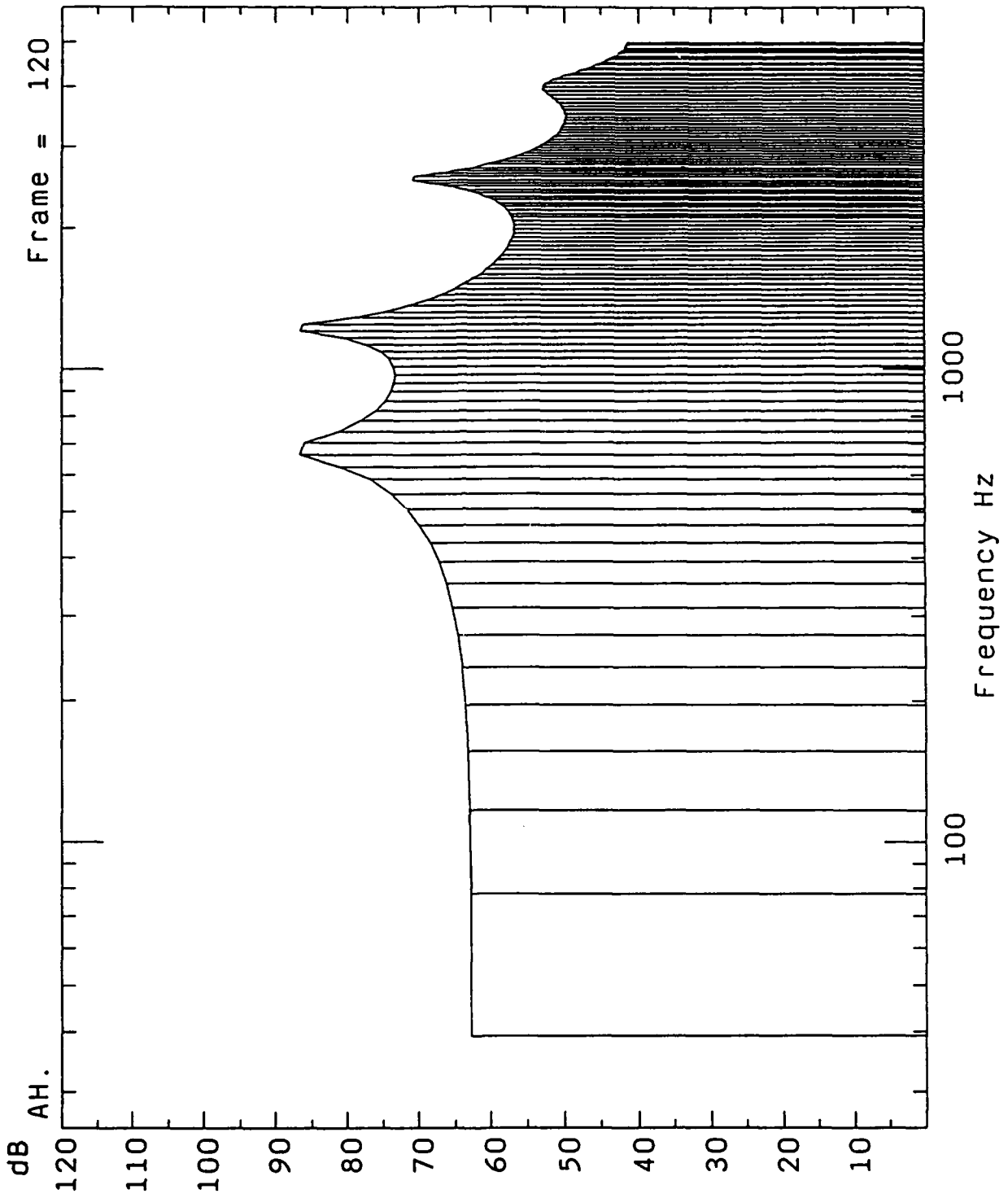


Figure C-8: Spectral envelope derived from FFT of [UH] reference token.

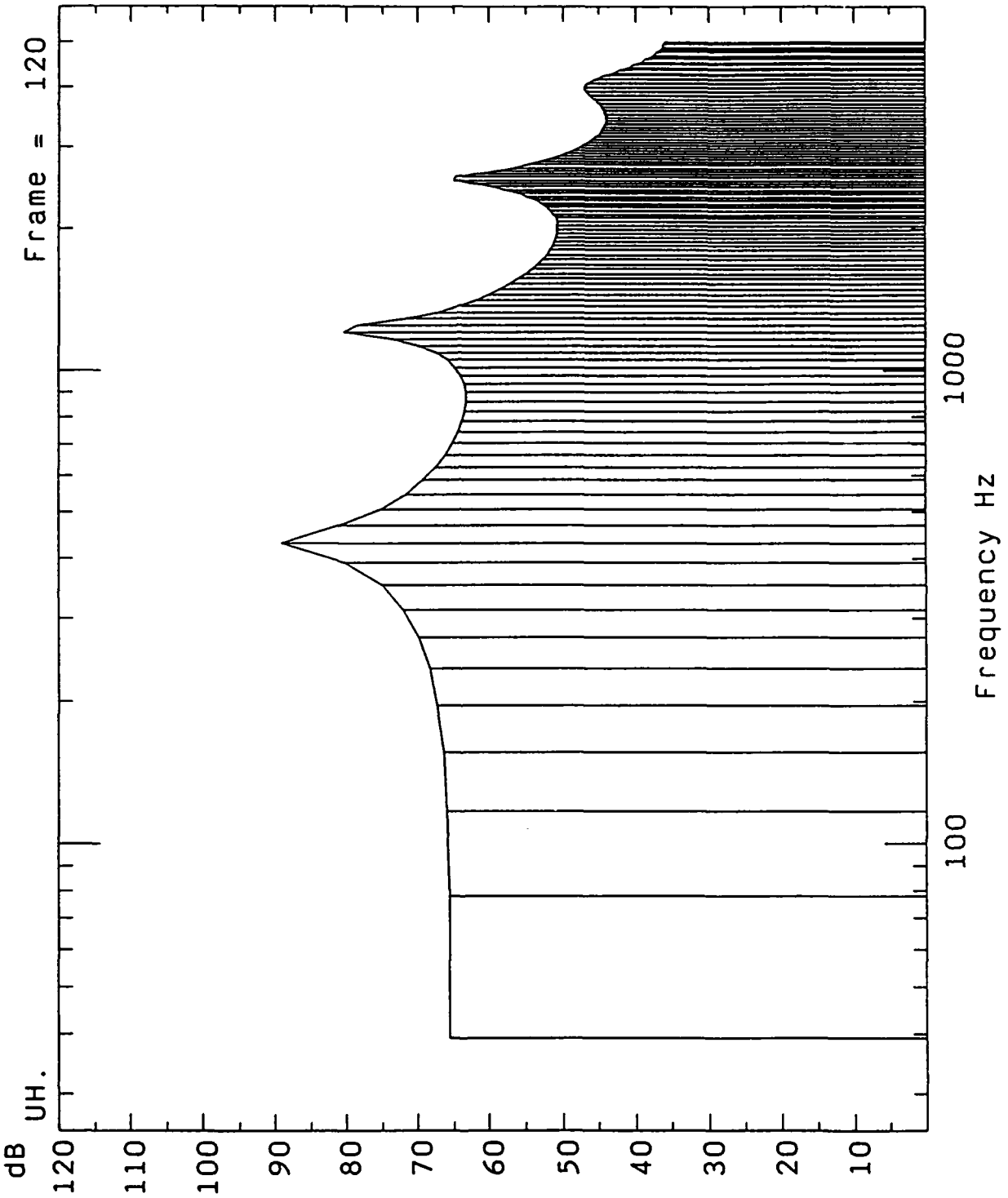


Figure C-9: Spectral envelope derived from FFT of [UW] reference token.

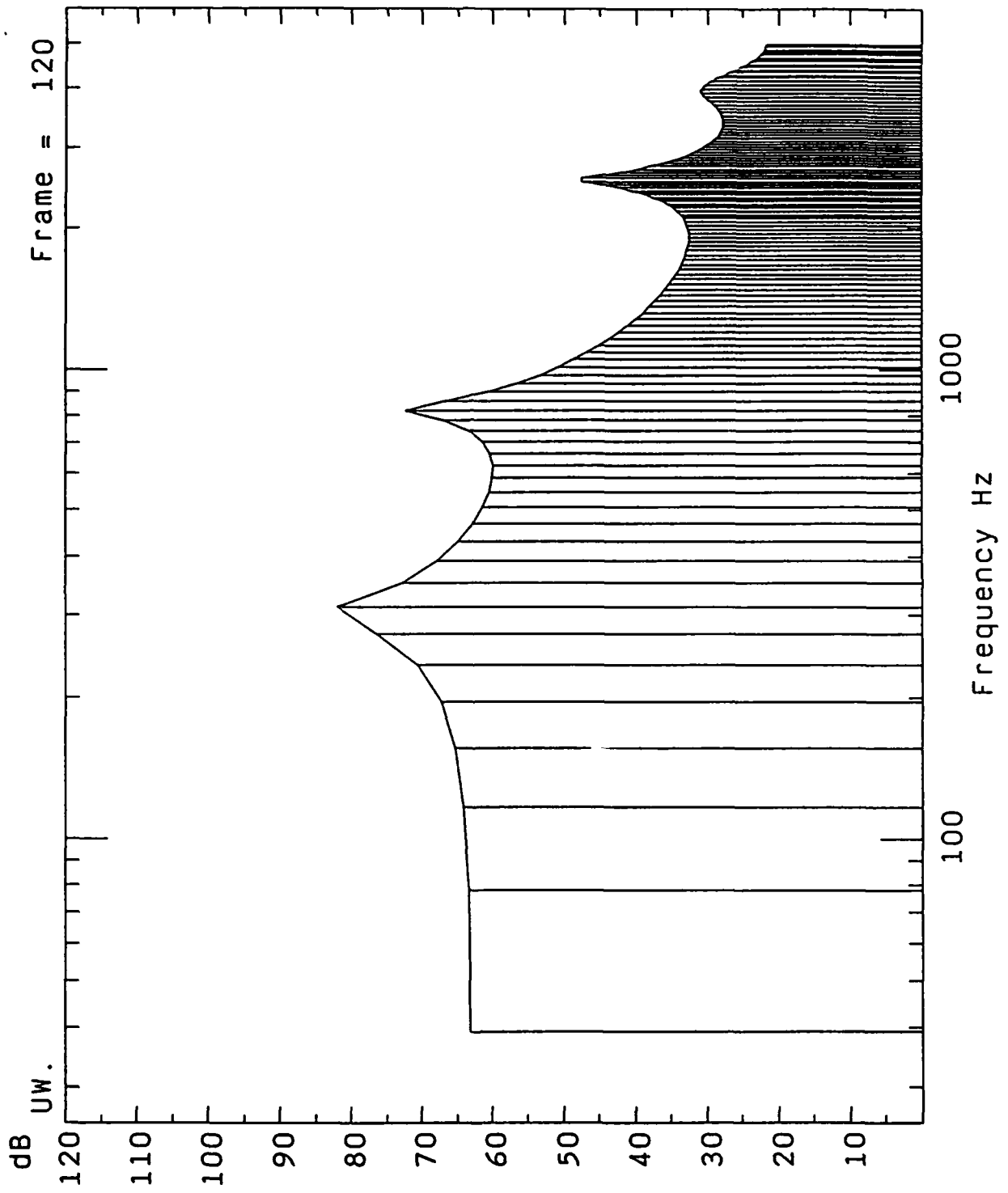


Figure C-10: Spectral envelope derived from FFT of [ER] reference token.

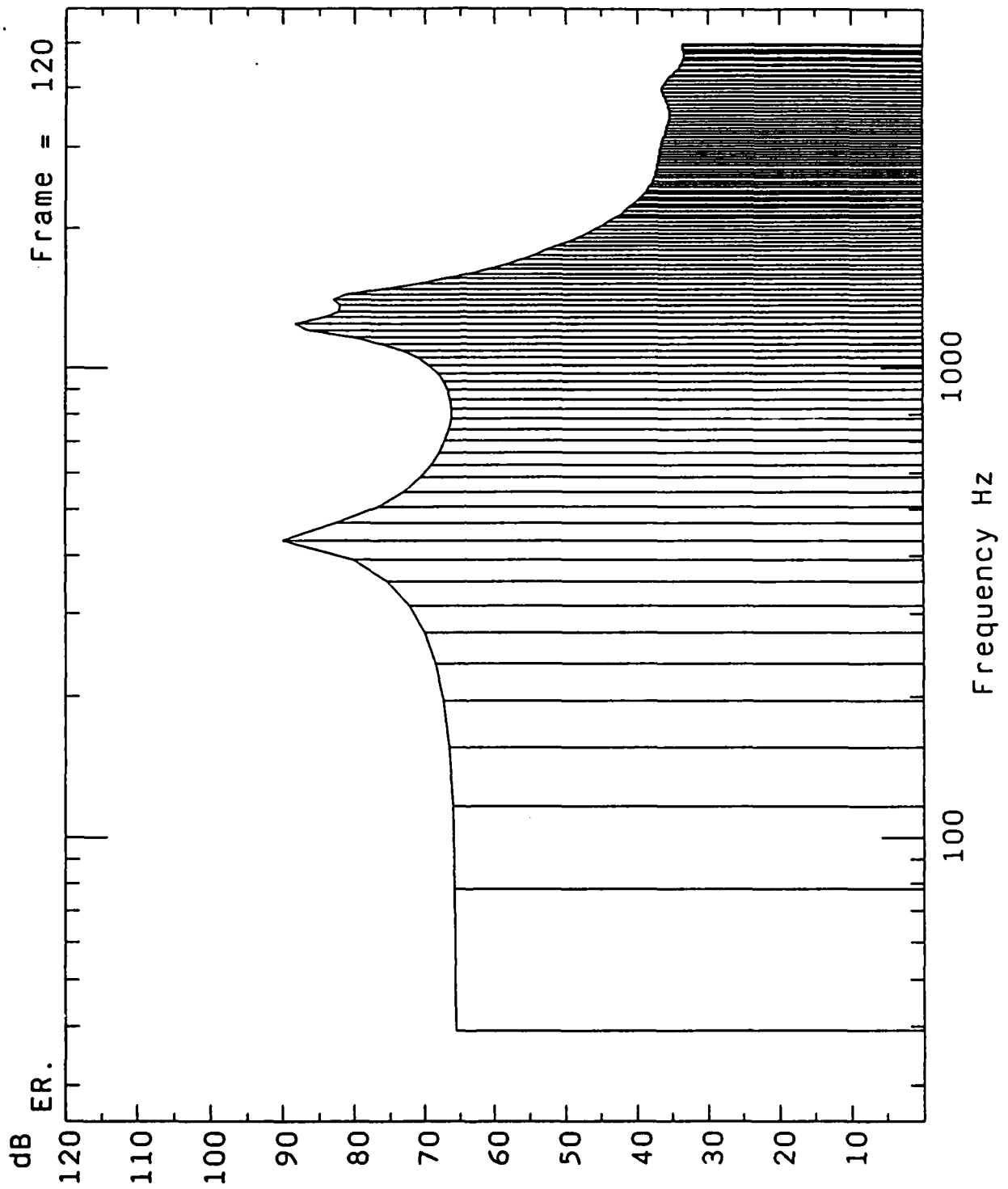


Figure C-11: Spectral envelope derived from FFT of [IY-IH] reference token.

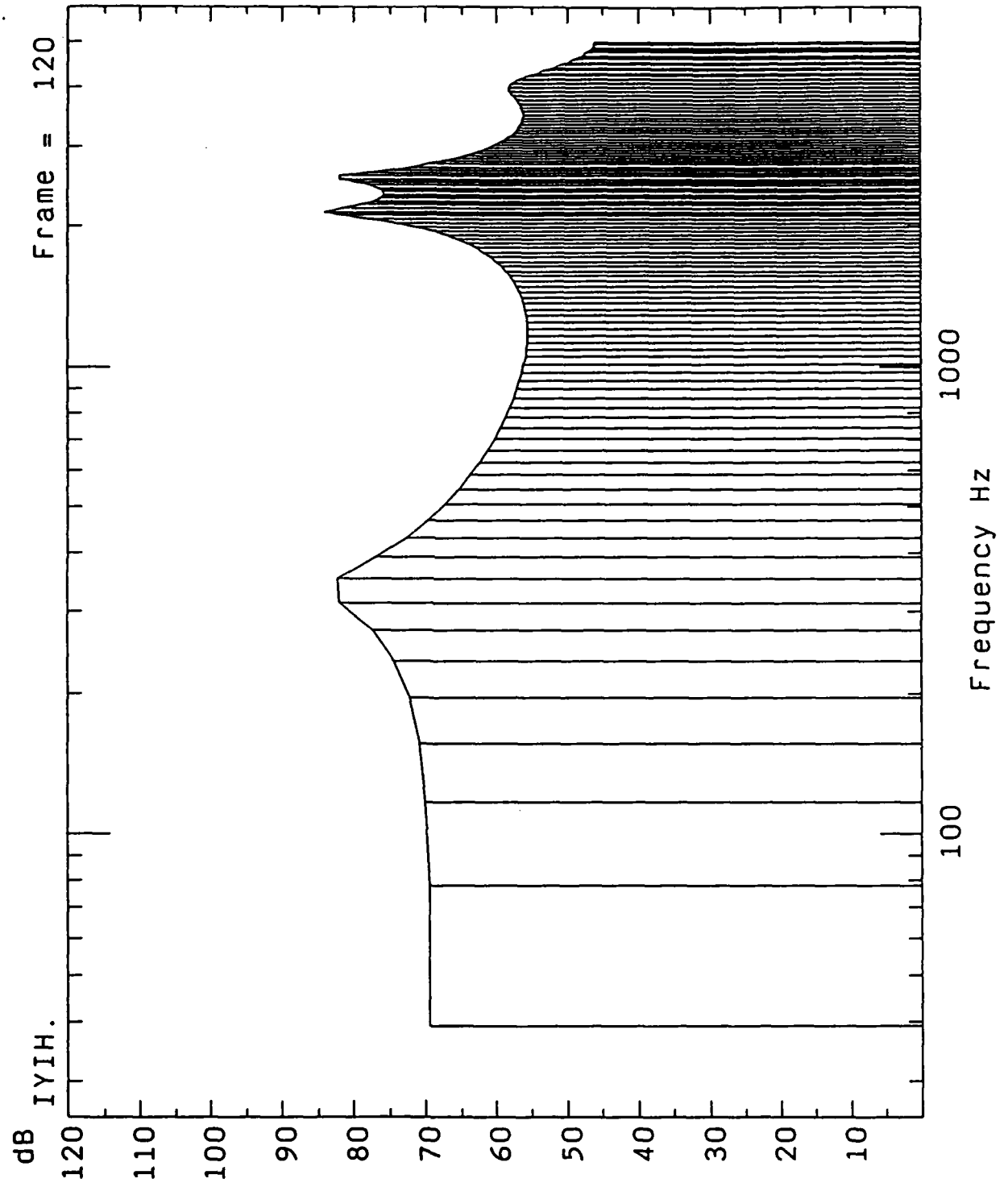


Figure C-12: Spectral envelope derived from FFT of [IH-EH] reference token.

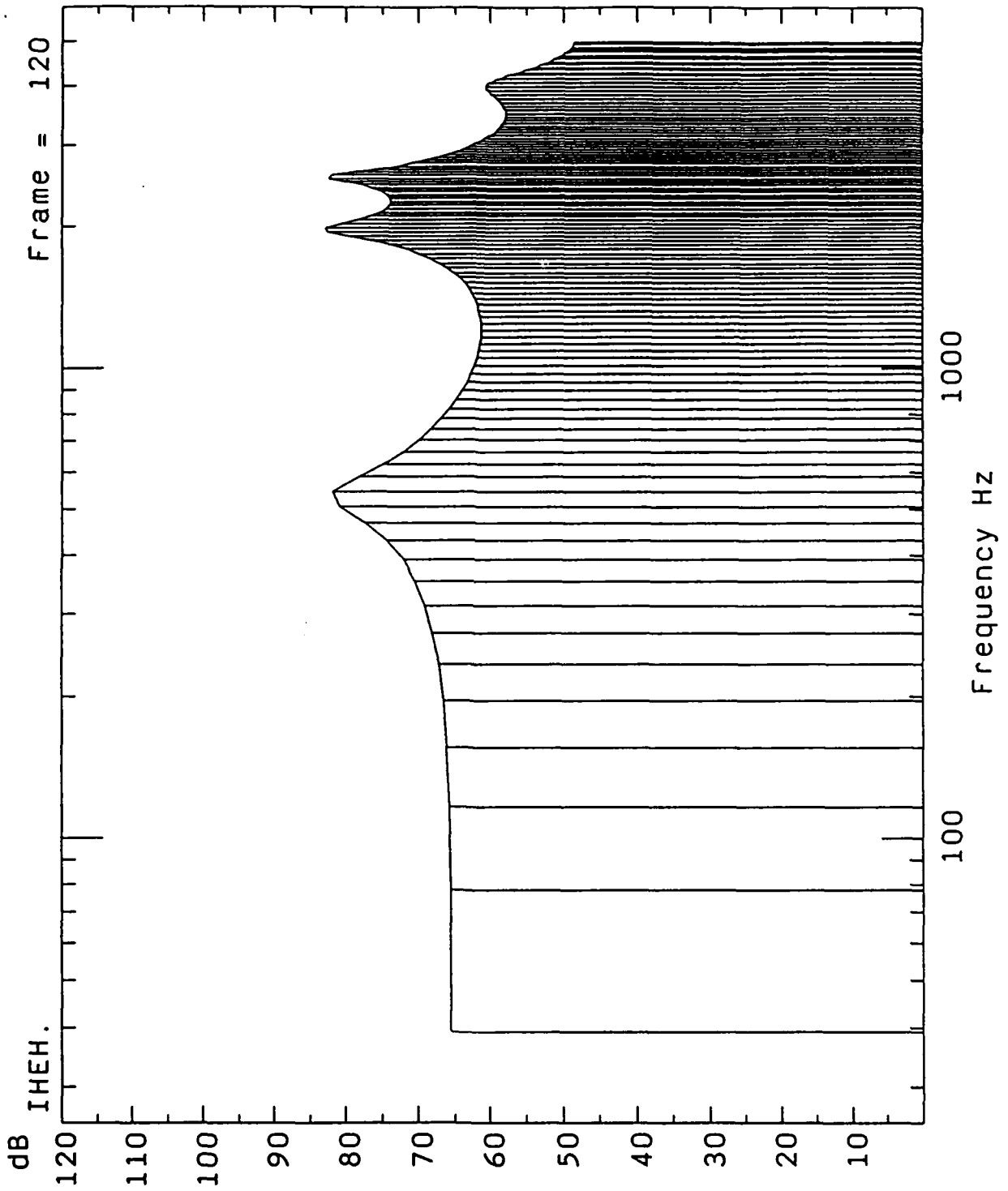


Figure C-13: Spectral envelope derived from FFT of [EH-AE] reference token.

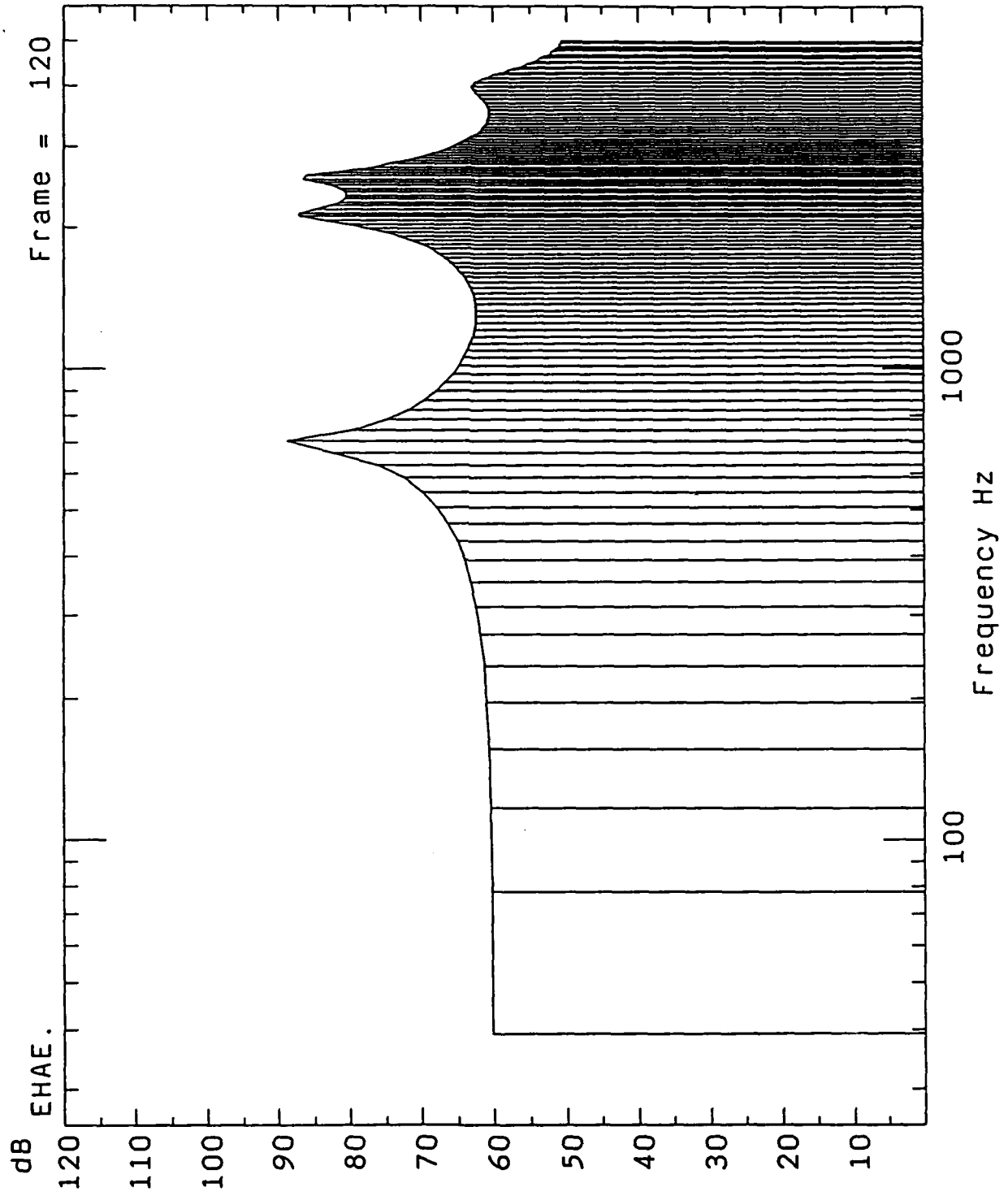


Figure C-14: Spectral envelope derived from FFT of [AE-AH] reference token.

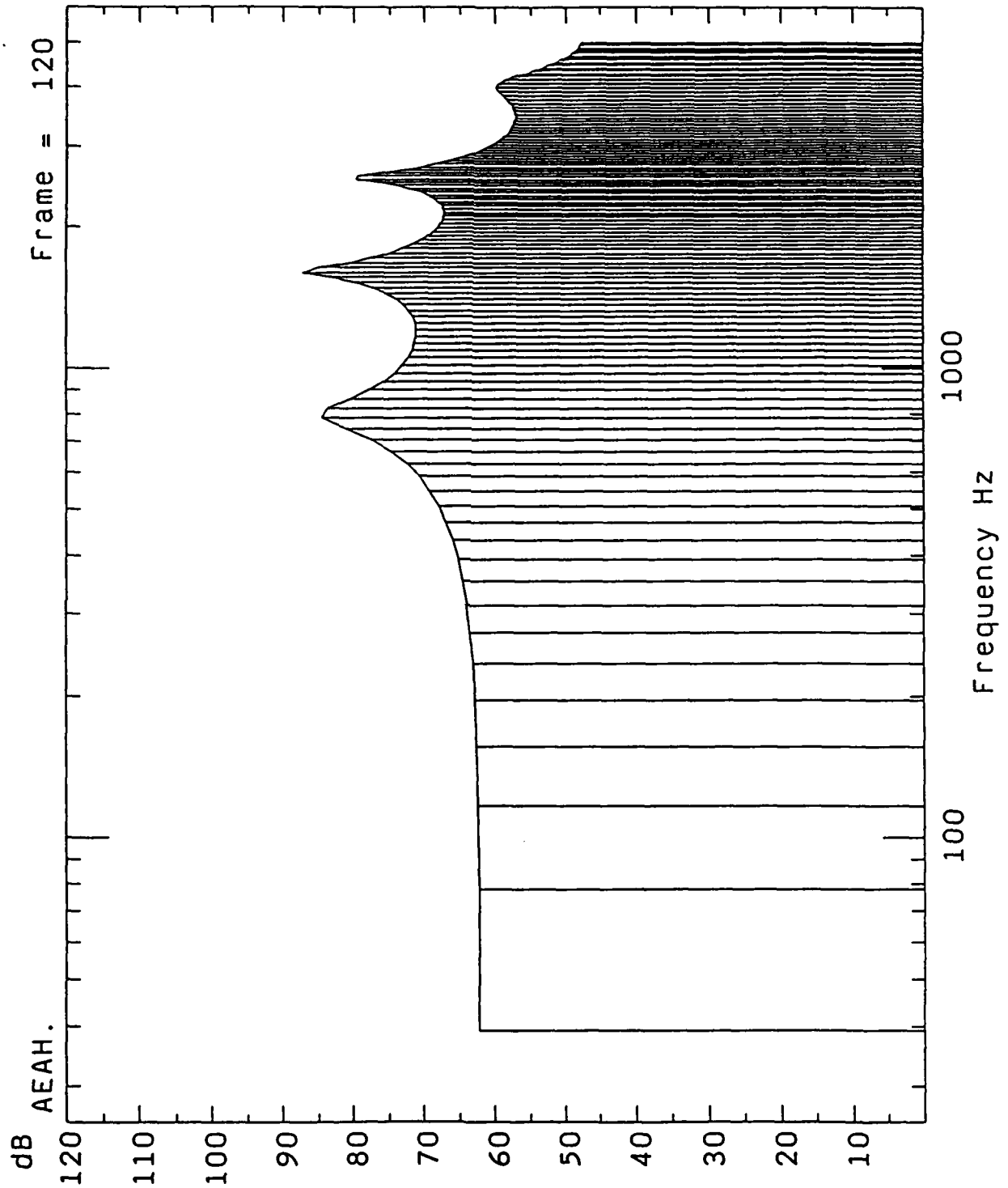


Figure C-15: Spectral envelope derived from FFT of [AH-AA] reference token.

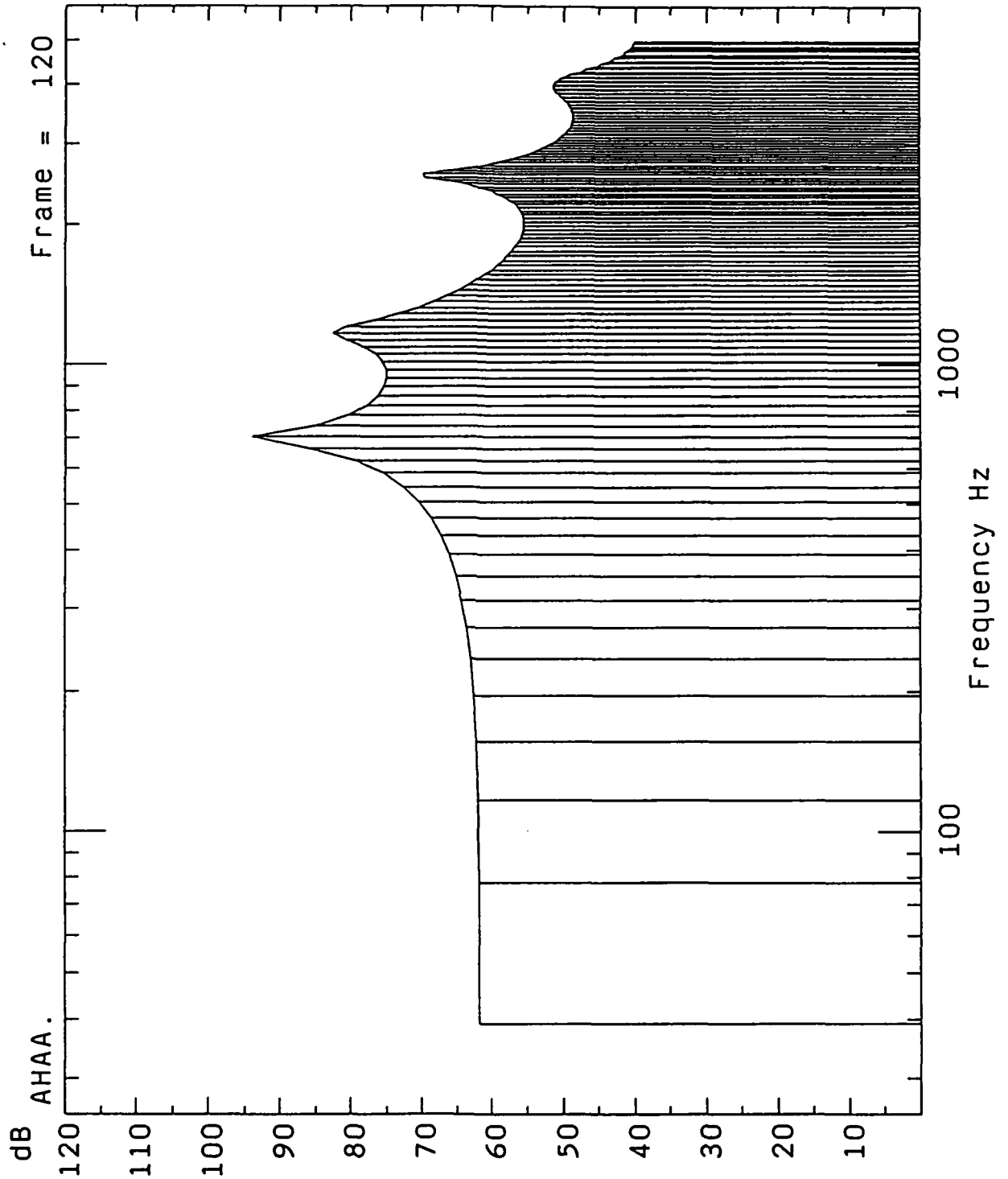


Figure C-16: Spectral envelope derived from FFT of [AH-UH] reference token.

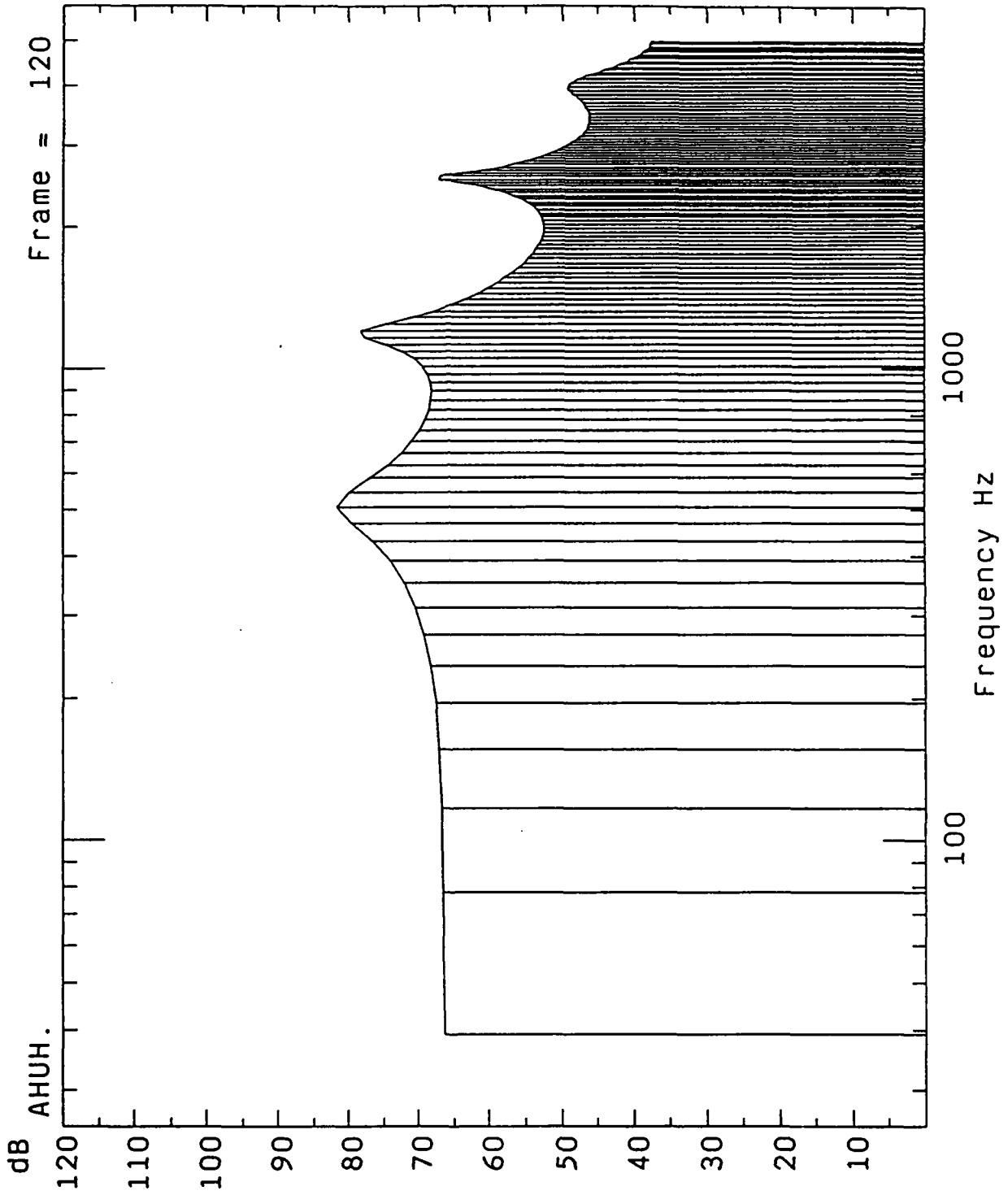
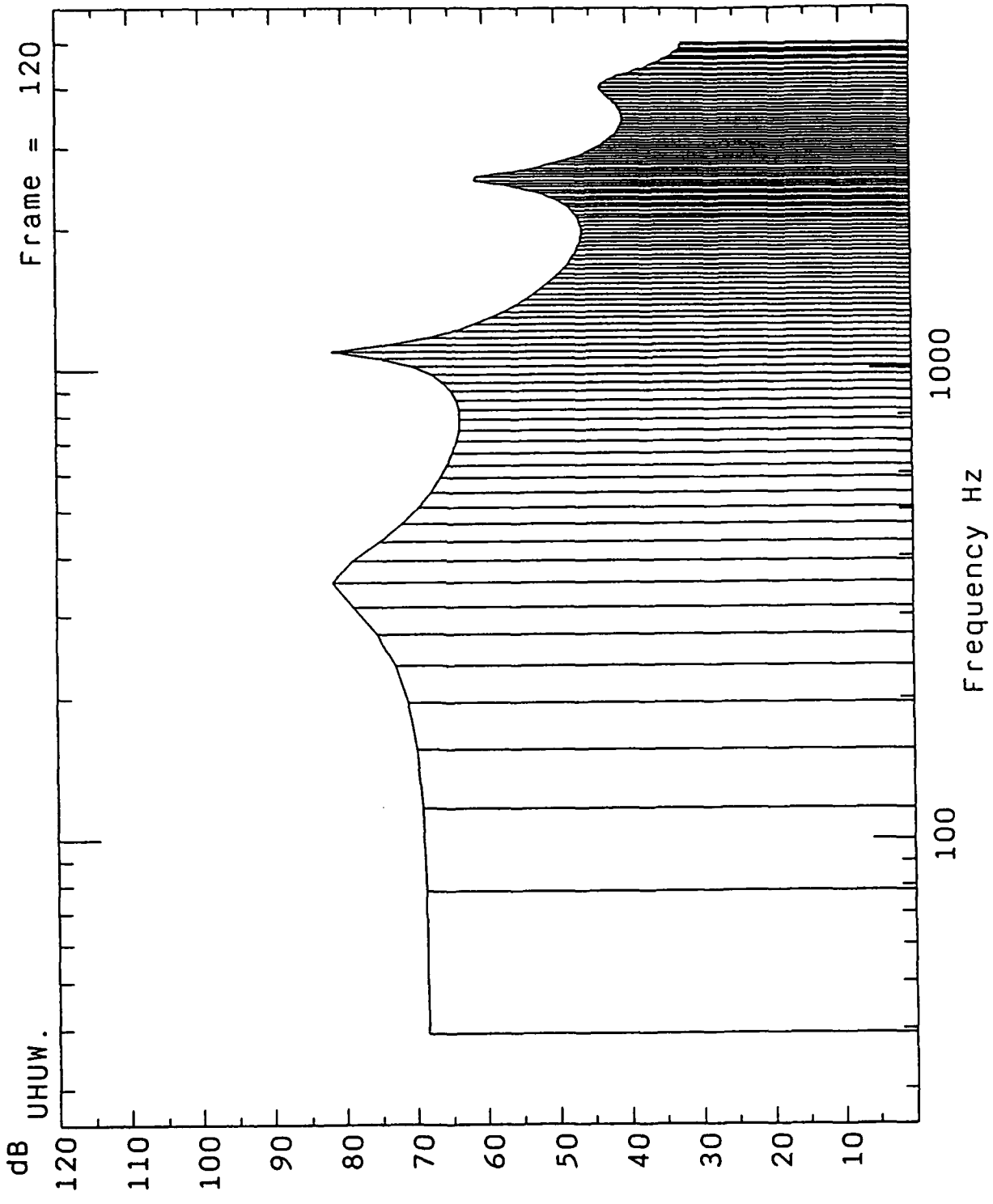


Figure C-17: Spectral envelope derived from FFT of [UH-UW] reference token.



Auditory-perceptual interpretation of the vowel

James D. Miller
Central Institute for the Deaf, St. Louis, Missouri 63110

AFOSR Grant 6-AFOSR-86-0335
Final Technical Report
Appendix

(Received 8 October 1987; accepted for publication 19 January 1989)

The major issues in relating acoustic waveforms of spoken vowels to perceived vowel categories are presented and discussed in terms of the author's auditory-perceptual theory of phonetic recognition. A brief historical review of formant-ratio theory is presented, as well as an analysis of frequency scales that have been proposed for description of the vowel. It is illustrated that the monophthongal vowel sounds of American English can be represented as clustered in perceptual target zones within a three-dimensional auditory-perceptual space (APS), and it is shown that preliminary versions of these target zones segregate a corpus of vowels of American English with 93% accuracy. Furthermore, it is shown that the nonretroflex vowels of American English fall within a narrow slab within the APS, with spread vowels near the front of this slab and rounded vowels near the back. Retroflex vowels fall in a distinct region behind the vowel slab. Descriptions of the vowels within the APS are shown to be correlated with their descriptions in terms of dimensions of articulation and timbre. Additionally, issues related to talker normalization, coarticulation effects, segmentation, pitch, transposition, and diphthongization are discussed.

PACS numbers: 43.71.An, 43.71.Es, 43.70.Fq

INTRODUCTION

The purpose of this paper is to describe a quantitative, theoretical approach to vowel perception. This approach is based on the auditory-perceptual theory of phonetic recognition (Miller, 1984a,b, 1987a,b), and it is designed to have the potential to provide a comprehensive and coherent account of the facts relating acoustic waveforms to perceived vowel categories. It is hoped that this theoretical approach to vowel perception will be useful not only as a guide to solutions of the unresolved problems in this area of investigation, but may also be useful in organizing and summarizing existing knowledge as well as integrating that knowledge into a general explanation of phonetic recognition.

Presently, it is well established that the locations of the first three prominences of the short-term spectrum of the vowel waveform are highly correlated with the perceived color of the vowel. However, this correlation is known to be attenuated by a variety of factors even when consideration is constrained to the carefully produced speech of a single dialect. Included among these attenuating factors are: differences between talkers' vocal characteristics associated with size, age, and gender; differences between talkers in articulatory style or habit; and differences in the acoustic expression of a vowel associated with large coarticulation effects induced by surrounding segments, with effects of speaking rate and linguistic stress, and with changes in, or modulations of, voice pitch. Additionally, when a vowel is uttered as part of a syllable or word, the sequence of short-term spectra often exhibits more than one segment that is nearly steady state, and sometimes it exhibits a nearly continuous change. There is no precise information as to which of these spectral patterns or which combination of these spectral patterns comprises the acoustic correlates of the perceived vowel. This is the problem of segmentation, and it also disturbs the measured correlation between spectral patterns and perceived vowels.

All of the factors mentioned above apply to the special case where the speaker and listener are both native speakers of the same dialect of their language, where the speech is accurately produced, where the semantics and syntax of the utterance are well known to the listener, and where the listener does not need to process extraneous noise or speech. In this special case, the issues of top-down cognitive and linguistic processing are largely removed from consideration. Even so, unresolved problems related to talkers, coarticulation, rate and stress, voice pitch, and segmentation remain, and each of these issues is addressed by the auditory-perceptual theory. It is claimed that these unresolved problems can either be resolved by the theory or that the theory provides an investigative framework that will lead to their resolution.

Since the auditory-perceptual theory has descended from the formant-ratio theory of vowel quality, and since it relies heavily on the logarithmic frequency scale, this paper begins with sections related to these issues. In Sec. I, a brief history of the formant-ratio theory is given, along with a summary of its advantages and shortcomings. Next, in Sec. II, frequency scales that have been suggested for the description of vowels are compared and discussed. Included in the discussion are mel, Bark, Koenig, log frequency, and frequency scales. Section III presents a brief description of the auditory-perceptual theory of phonetic recognition and demonstrates how it can be applied to a sustained vowel and a consonant–vowel syllable. Concepts to be described include the sensory reference, sensory formants, the sensory-perceptual transformation, perceptual formants, the auditory-perceptual space, perceptual target zones, and segmentation maneuvers. These concepts are illustrated by data gleaned from the literature and by measures made in

 0001-4966/89/052114-21$00.80 © 1989 Acoustical Society of America

our laboratories at Central Institute for the Deaf (CID). Also, included in Sec. III is a description of preliminary estimates of the perceptual target zones for the simple (that is, monophthongal) vowels of American English. In Sec. IV, characterizations of the vowels in the auditory-perceptual space are shown to be related to descriptions in terms of articulation and timbre. Section V deals with four separate issues in vowel perception: (a) voice pitch as a determiner of vowel category, (b) transposition of vowel spectra, (c) vowels with changing formants, and (d) lability of vowel boundaries. Section VI contains concluding comments.

I. FORMANT-RATIO THEORY

In the late 1800s, Richard John Lloyd (1890a,b; 1891; 1892) formulated the formant-ratio theory of the vowel. Lloyd's dictum was that like articulations produce like perceptions of vowel qualities, and that like articulations produce like ratios of the formants. He called his theory "the relative resonance theory" and stated that vowel quality depends on the intervals between the resonances, not their absolute values. Lloyd based his theory on the central role of the frequency ratio in music, on acoustic differences in the vowels of men, women, and children, on experiments with synthetic vowels produced by exciting hand-blown glass tubes, and on studies of transposition done through synthesis or by speeding and slowing the playback of Ediphone recordings. His critics [for example, Pipping (1893)] were quick to respond by pointing out that certain distinct vowels have common ratios of second and first resonant frequencies, that is, $F2/F1$ ratios, that transposition works only over certain limited ranges, that some vowels seem to have only a single formant rather than two, and that his formulas for calculating the resonances of his tubes were in error. In reply, Lloyd (1894) asserted that almost all vowels were composed of at least two resonances and that errors in calculating resonances were irrelevant to his thesis that like articulations produce like formant ratios and like vowel percepts. He defended his experiments in transposition as showing that vowel perception was much more sensitive to changes in formant ratios than to changes in absolute values that maintain appropriate ratios when both kinds of change are measured in semitones. But he did admit that, while the formant ratio was the dominant factor in vowel perception, absolute frequencies must also play a role.

Statements of the formant-ratio theory appear in the literature every few years since Lloyd's work (as examples, see Chiba and Kajiyama, 1941; Potter and Steinberg, 1950; Iri, 1959; Peterson, 1961; Okamura, 1966; Minifie, 1973; Broad, 1976; and Kent, 1979), and, interestingly, the authors usually seem to be unaware of prior descriptions of the notion. Among these papers, Peterson's (1961) discussion is particularly useful and penetrating. The major strength of formant-ratio theory is its ability to eliminate or dramatically reduce differences in the acoustic description of vowels related to talker, age, and gender. The major stumbling blocks to the acceptance of formant-ratio theory seem to be related to failures to provide clear answers to the following questions. Why does the transposition of vowels hold only over certain restricted ranges on the log frequency scale? Why do some differently perceived vowels have similar formant ratios? And, how is it that different instances of the same vowel can have very different formant ratios? The strengths and weaknesses of formant-ratio theory as applied to natural speech are beautifully illustrated in the quantitative work of Potter and Steinberg (1950), Peterson and Barney (1952), and Peterson (1961).

Peterson and Barney (1952) plotted $F2$ values against $F1$ values, as shown in Fig. 1 (see Ref.1). On this figure, frequency is plotted in accordance with the Koenig scale (Koenig, 1949). The lines corresponding to equal $F2/F1$ ratios have been added to clarify this presentation. It can be seen that, while the data for a given vowel tend to cluster around a line representing a given $F2/F1$ ratio, there is considerable scatter. Also, within each of the vowel groups [/AA/ and /AO/], [/UH/, /UW/, and /AE/], and [/UW/, /ER/, and /EH/], the $F2/F1$ ratios are very similar. Earlier, Potter and Steinberg (1950) made similar observations on the simple vowels of American English as spoken by ten men, ten women, and five children. As shown in Fig. 2, they noted strong correlations between the pitch and formant values within each vowel category. They also noted, as shown in Fig. 3, that characterizing the vowels by ratios of the mel equivalents of $F1$, $F2$, and $F3$, that is, by $M3/M2$ and $M2/M1$, almost eliminated talker differences within each vowel category. Also, they stated that ratios of formant frequencies behaved exactly the same as mel ratios by way of eliminating talker differences. Thus Potter and Steinberg in 1950 once again offered strong evidence in support of Lloyd's view that like formant ratios result in like perceived

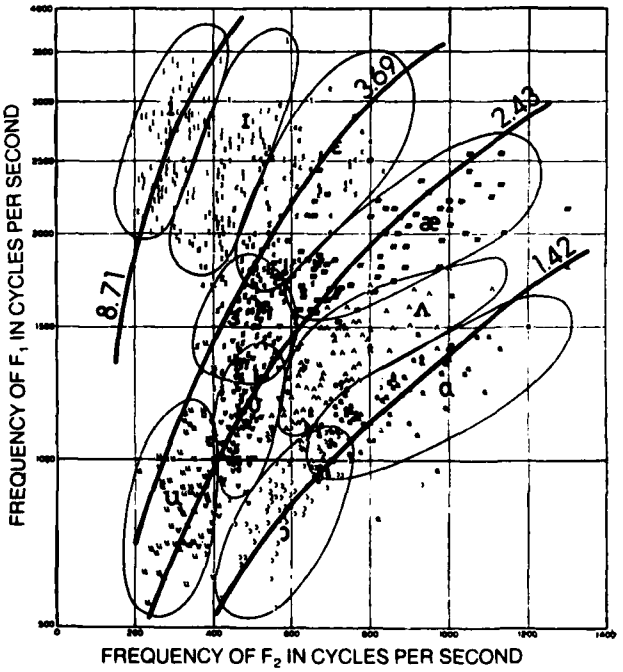


FIG. 1. Scatter of vowels in $F2$ by $F1$ plot where both variables are on the Koenig scale. Lines representing typical values of $F2/F1$ for the vowels /IY/, /EH/, /AE/, and /AA/ are shown. Note that within each of the vowel groups [/AA/ and /AO/], [/UW/, /UH/, and /AE/], and [/UW/, /ER/, and /EH/], the $F2/F1$ ratios can be very similar (after Fig. 8 from Peterson and Barney, 1952).

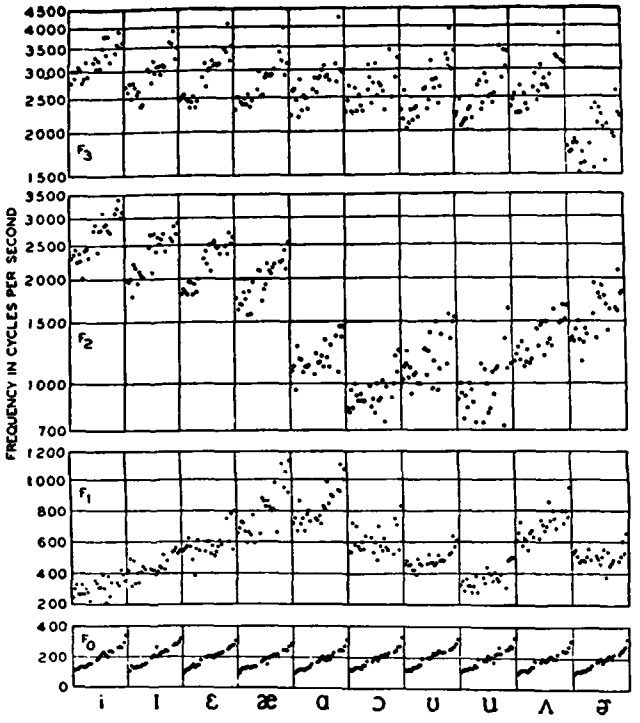


FIG. 2. The pitch and formant values of the simple vowels of American English. On each panel from left to right are the data for ten men, ten women, and five children. Within each group of speakers, the data are arranged from left to right in order of increasing pitch. This is a reprint of Fig. 11 from Potter and Steinberg (1950).

vowels. However, Potter and Steinberg were discouraged by the fact that neither [/AA/ and /AO/] nor [/UH/ and /UW/] could be distinguished by formant-ratio theory. In these cases, some other factor, perhaps akin to the absolute locations of the formants along the frequency scale, must play a role.

Later, Peterson (1961) returned to this issue. He confirmed the observations of Potter and Steinberg by showing that ratios of the center frequencies of the first and second formants and of the second and third formants served to eliminate talker differences whether these frequencies are expressed in hertz or in mels. Peterson then demonstrated that mel ratios and frequency ratios are nearly identical over the range usually occupied by the first three formants of vowels and that the data fit either formulation nearly equally well, as shown in Fig. 4.

The early work of Lloyd, and the subsequent works cited above, all indicate that use of the ratios of the center frequencies of the first three formants can reduce and nearly eliminate talker differences. It is largely for this reason that such ratios play a prominent role in the auditory-perceptual interpretation of the vowel. However, additional concepts are introduced to deal with the shortcomings of formant-ratio theory mentioned above. Before presenting the auditory-perceptual theory, a more detailed discussion of frequency scales follows in the next section. This section is meant to rationalize the use of the log frequency scale in the auditory-perceptual theory and to clarify the relations of the log frequency scale to its prominent alternatives.

II. FREQUENCY SCALES IN VOWEL ANALYSIS

Various scales of frequency have been suggested for vowel analysis. Most of these suggestions are based on the notion that auditory scales, such as the mel scale (Fant, 1973), the Bark scale (Zwicker and Terhardt, 1980), the Koenig scale (Koenig, 1949), or cochlear position scales (Greenwood, 1961), reflect auditory analysis more accurately than others. These scales, whether based on frequency-position maps of the cochlea, critical-band measures, or pitch scaling experiments, all tend to be linear functions of frequency in hertz in the low-frequency region, transitional in the mid-frequency region, and logarithmic in the high-frequency region. Consider Fig. 5, where normalized values of the mel, Bark, and Koenig scales are plotted against frequency (f) in hertz, and both the x and y scales are logarithmic. The mel values were calculated from Fant's (1973) equation for technical mels (TM),

$$TM = (1000/\log 2)\log(f/1000 + 1). \quad (1)$$

The Bark values (B) were calculated by the equation from Zwicker and Terhardt (1980),

$$B = 13 \arctan(0.76f/1000) + 3.5 \arctan(f/7500)^2. \quad (2)$$

The Koenig values (K) were calculated from the equations (Koenig, 1949)

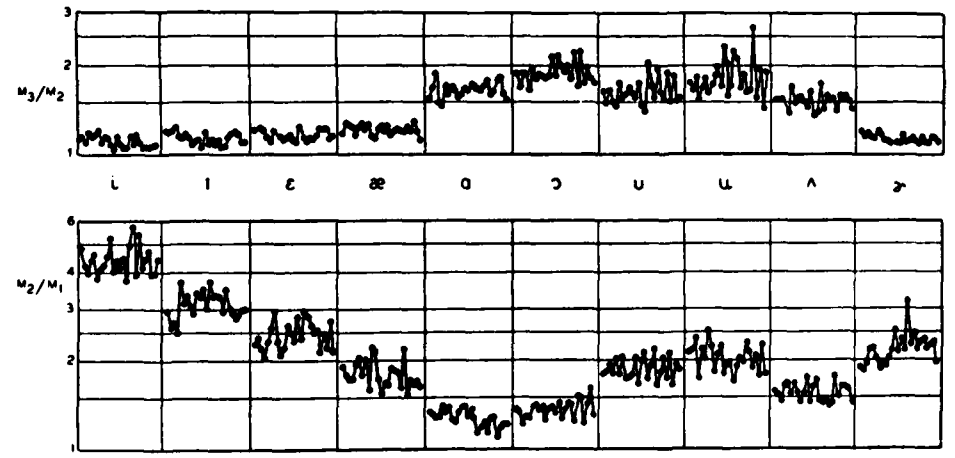


FIG. 3. The ratios of the mel equivalents of $F3$ and $F2$ ($M3/M2$) and of $F2$ and $F1$ ($M2/M1$) for the data shown in Fig. 2. Note that talker differences are removed, but that the pairs [/AA/ and /AO/] and [/UH/ and /UW/] cannot be distinguished. This is a reprint of Fig. 13 from Potter and Steinberg (1950).

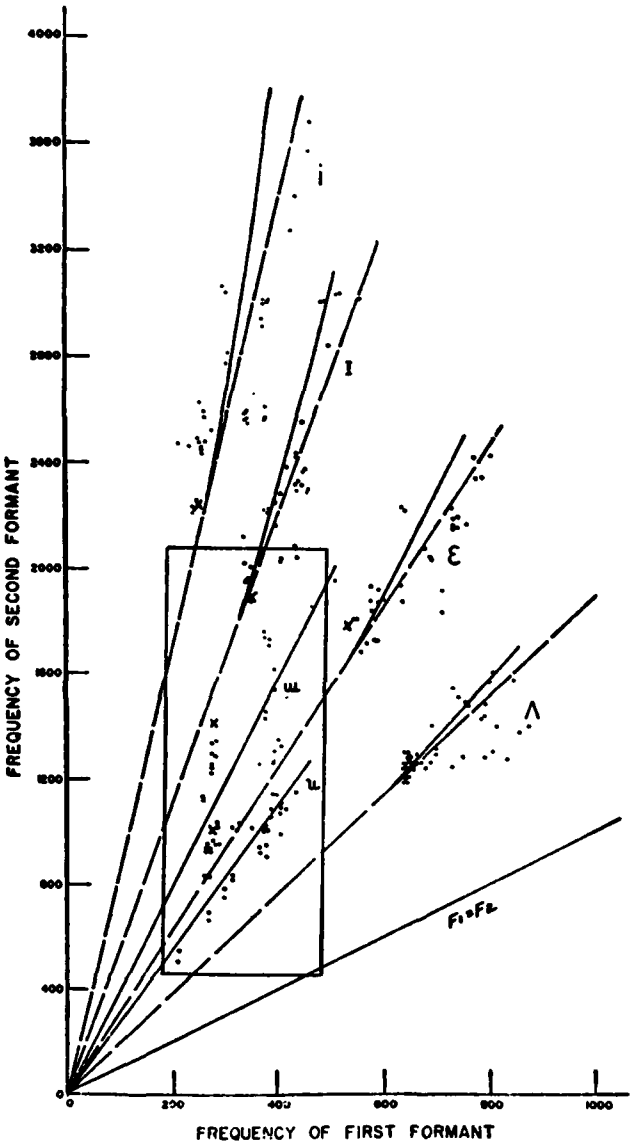


FIG. 4. The frequency of the first and second formants for a set of matched vowels by men, women, and children speakers. Points for a given vowel cluster close to constant frequency ratio (dashed) or constant mel ratio (solid) lines. (The large rectangle on the figure is not relevant to the present paper.) This is a reprint of Fig. 2 from Peterson (1961). (Figure reprinted with the permission of ASHA.)

$$K = 0.002f \quad \text{for } 0 \leqslant f \leqslant 1000, \tag{3}$$

$$K = (4.5 \log f) - 11.5 \quad \text{for } 1000 \leqslant f \leqslant 10\,000. \tag{4}$$

For purposes of graphic comparison, the calculated Bark values were multiplied by 117.5015 and the Koenig values were multiplied by 500 so that all scales would have a value of 1000 for the frequency of 1000 Hz. As can be seen in Fig. 5, all four scales have similar slopes up to about 1000 Hz, and from 1000–4000 Hz the Koenig, mel, and Bark scales make the transition to a more nearly logarithmic slope. Interestingly, the change in slope from the range of 100–1000 Hz to the range of 1000–4000 Hz is greatest for the Koenig scale, intermediate for the Bark scale, and least for the mel scale. Otherwise said, the mel scale is most nearly a linear function of frequency, while the Bark scale is intermediate, and the Koenig scale least in this regard. In spite of

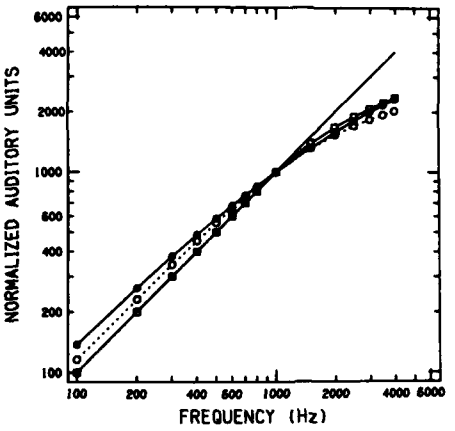


FIG. 5. The mel (closed circles), Bark (open circles), Koenig (open squares), and frequency (solid line) scales are plotted against frequency. The Bark values are multiplied by 117.5015 and the Koenig values are multiplied by 500 so that all scales have a value of 1000 for the frequency of 1000 Hz. Note the similarity of all four scales.

these deviations from linearity, Fig. 5 supports the idea that, over most of the ranges of values of the center frequencies of $F1$, $F2$, and $F3$, the Koenig scale, the mel scale, the Bark scale, and frequency in hertz are nearly equivalent.

A corollary of this observation is that formant ratios, whether in mels, Barks, Koenigs, or hertz, should be nearly equivalent over the ranges associated with the vowels of American English. To test this hypothesis, we calculated these ratios for the mean data of Peterson and Barney (1952) for nonretroflex vowels. Since there are nine vowels and three talker groups, there were 27 $F2/F1$ ratios and 27 $F3/F2$ ratios. The results of these calculations are shown in Fig. 6. In panel (a), the logs of the ratios of the mel values of the formants are plotted against the logs of the frequency ratios. An extremely high correlation is evident, and the conclusion of Potter and Steinberg (1950) and of Peterson (1961) that mel ratios are equivalent to frequency ratios in describing vowels is confirmed. Similar, but less precise, correlations are noted for the Bark transformation in panel (b) and the Koenig transformation in panel (c). These results reflect the deviations from linearity noted in the discussion of Fig. 5; that is, the greater the deviation from linearity noted on Fig. 5, the poorer the correlation noted in Fig. 6.

In contrast with the use of ratios or the logs of ratios, it has sometimes been suggested that differences in the variables describing $F1$, $F2$, and $F3$ might be an effective metric for vowel quality. For example, Syrdal and Gopal (1986) have shown that certain feature descriptions of vowels can be predicted by particular values of differences in the Bark values of the center frequencies of the formants. Figure 7 exhibits the correlations between such differences in mels [panel (a)], in Barks [panel (b)], and in Koenigs [panel (c)]. Here, it can be seen that the correlations with the logs of the frequency ratios are best for differences in Barks and Koenigs and a bit poorer for differences in mels. The correlations between logs of frequency ratios with Bark differences or with Koenig differences would be nearly perfect except for a cluster of points associated with the $F1$ and $F2$ values of the vowels /UW/ and /UH/. Figure 8 is included for complete-

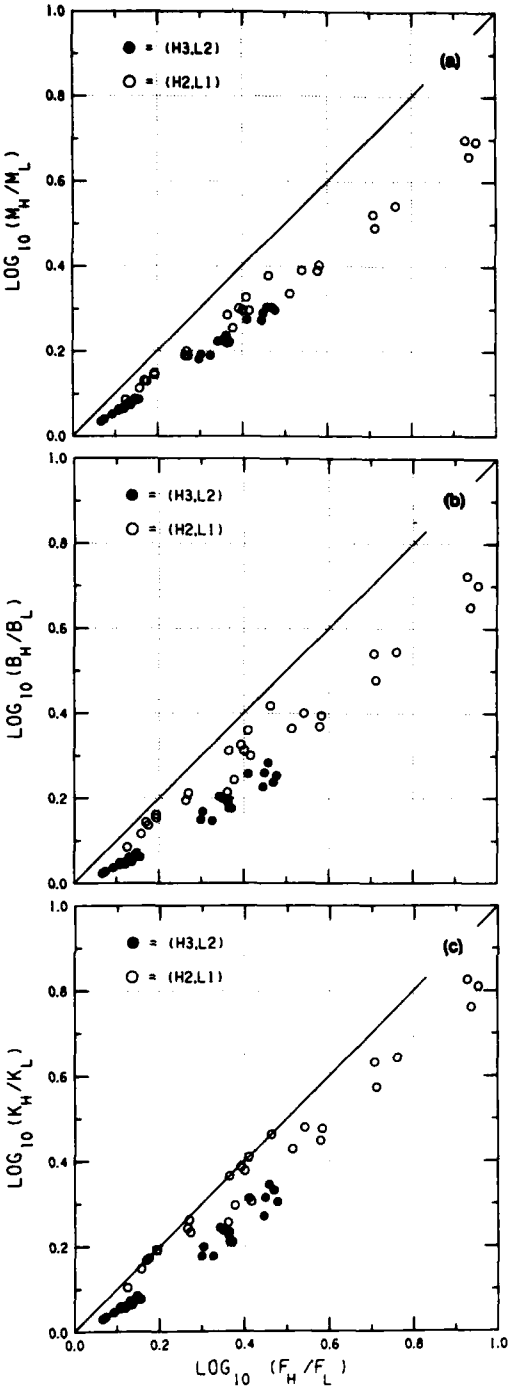


FIG. 6. Scatter plots are shown for the relations between logarithms of formant ratios expressed in hertz [x axes of panels (a), (b), and (c)] and logarithms of formant ratios expressed in mels [y axis of panel (a)], Barks [y axis of panel (b)], and Koenigs [y axis of panel (c)]. The data are the mean values of $F1$, $F2$, and $F3$ given by Peterson and Barney (1952) for nine nonretroflex vowels of American English for three talker groups. Thus there are 27 values of the ratio $F2/F1$ (open circles) and there are 27 values of the ratio $F3/F2$ (closed circles) on each plot. The $F2/F1$ ratios for various scales are indicated as ($H2/L1$), while the $F3/F2$ ratios are indicated as ($H3/L2$).

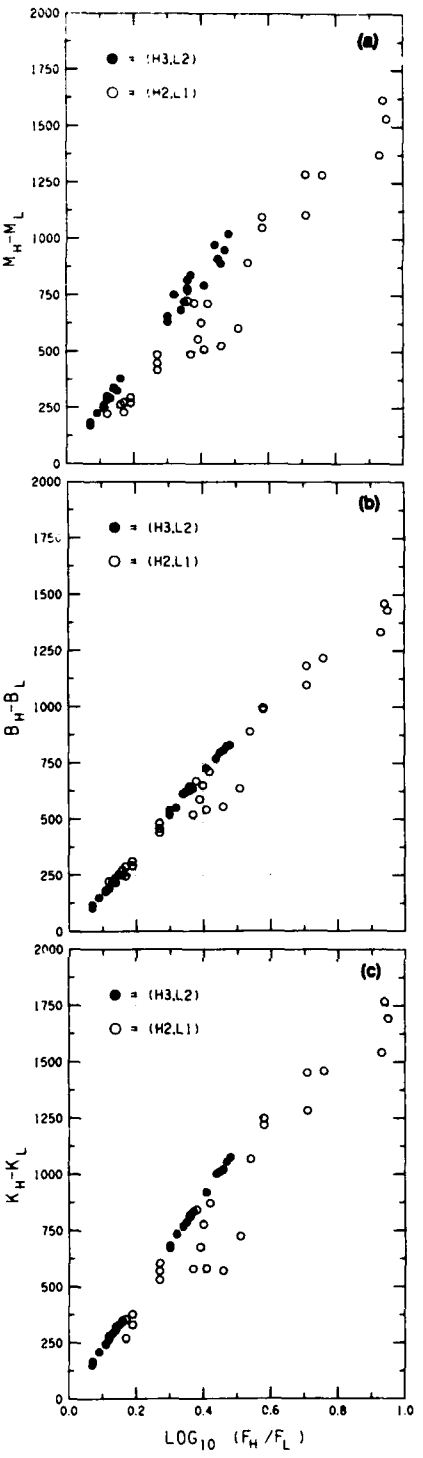


FIG. 7. Scatter plots are shown for the relation between logarithms of formant ratios expressed in hertz [x axes of panels (a), (b), and (c)] and differences in formant values expressed in mels [panel (a)], in Barks × 117.5015 [panel (b)], and in Koenigs × 500 [panel (c)]. Open circles are for $F2 - F1$ differences and closed circles are for $F3 - F2$ differences. Data are the same as for Fig. 6.

ness, and it demonstrates that differences in formant values in hertz do not correlate as well with the logs of the frequency ratios as do the differences in mels, Barks, or Koenigs shown in the previous figure.

Examination of Figs. 5–8 suggests that use of differences, ratios, or log ratios of any of the suggested scales might be about equally effective in clustering the vowels independent of talker group. The degree of clustering for the

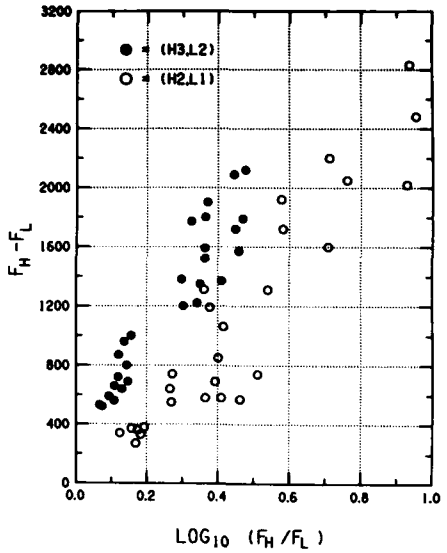


FIG. 8. A scatter plot is shown for the relation between logarithms of formant ratios and differences in formant values. Both scales are in hertz. Symbols and data are the same as in Fig. 7.

log ratios and differences of ($F1, F2$) and ($F2, F3$) is shown for the four scales in Fig. 9. On each panel, the vowel is indicated by an Arabic number as coded in the legend. By visual inspection, one can see that all of the ratio descriptions cause the like vowels to cluster well, with fair to good separation of unlike vowels. Differences also serve to cluster the vowels. In the cases of the mel, Koenig, and frequency scales, the log ratio approach seems to do a better job of clustering the vowels than does the difference approach. In the case of the Bark transformation, the two approaches appear to be nearly equivalent.

A mathematical analysis confirms the visual impressions of Fig. 9. Following the work of Miller *et al.* (1983), we calculated the distances between every pair of points in a plane. Next, we calculated the mean of the squares of all distances between vowel categories (σ^2_{bcd}) and the mean of the squares of all distances within vowel categories (σ^2_{wcd}). The square root of the ratio of these mean squares, $(\sigma^2_{bcd}/\sigma^2_{wcd})^{1/2}$, is the normalized root-mean-square distance between vowel categories and is symbolized by d'_{rms}. The larger this number, the better the overall clustering by vowel category. Unfortunately, d'_{rms} can be misleading in the following way. Several vowel categories could overlap completely, but d'_{rms} could yet be large if one or more categories were greatly separated from the others. Therefore, a second statistic was calculated. The mean position of each vowel in the plane under consideration was calculated, and the *smallest* distance between means (d_{mn}) was found. Then, the statistic $d'_{mn} = d_{mn}/\sigma_{wcd}$ was calculated representing the minimum separation between vowel categories expressed in standard deviation units.

Using the statistics just described, we ranked the vowel metrics discussed above with regard to their effectiveness in clustering the vowel data shown in Fig. 9. The results are shown in Table I. While there are slight differences in the

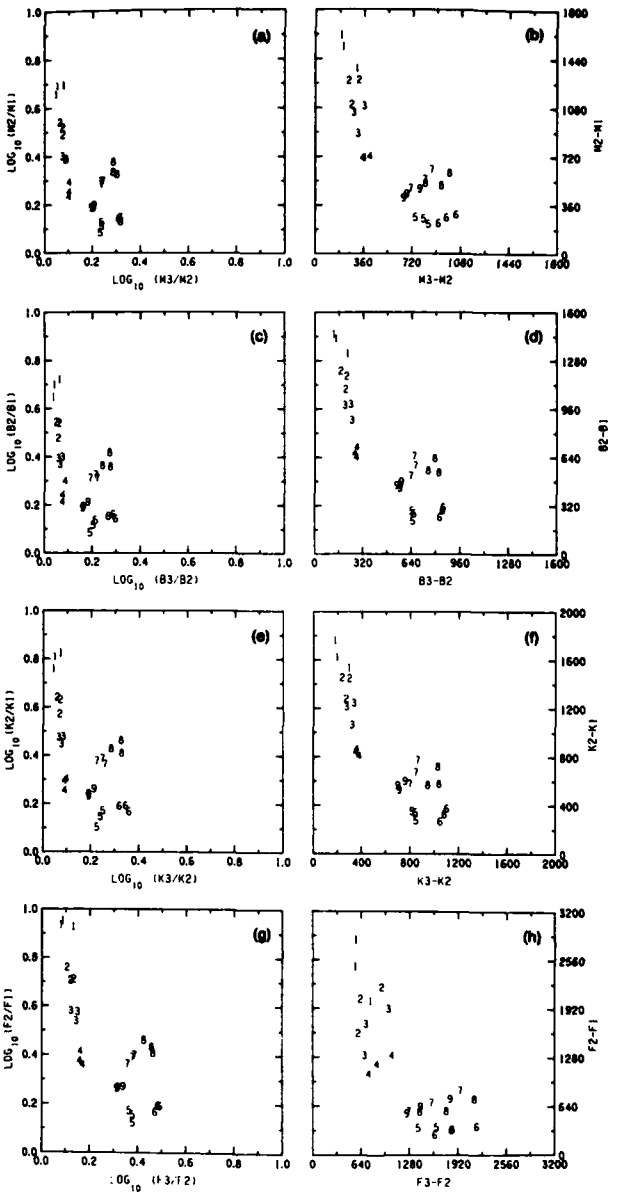


FIG. 9. The clustering of vowels is shown for various expressions of the relations among $F1$, $F2$, and $F3$. The vowels are indicated by Arabic numbers: 1—/IY/; 2—/IH/; 3—/EH/; 4—/AE/; 5—/AA/; 6—/AO/; 7—/UH/; 8—/UW/; and 9—/AH/. One value for each vowel is shown for mean data of men, women, and children (Peterson and Barney, 1952). The reader should note the spacing *within* vowel categories in relation to that *between* vowel categories.

rankings produced by the two statistics, a clear ordering emerges.

When the vowels are specified by the logs of the ratios of formant center frequencies measured in hertz or mels, the best clustering is observed. Clearly, inferior clustering is observed for differences of the center frequencies of the formants in Koenigs, mels, or hertz. Intermediate clustering is observed when logs of Koenig ratios, Bark differences, or logs of Bark ratios are used as descriptors of the relations among the formants.

The author's preference for the log frequency scale is based on the literature and the results described above, on the results of similar, more extensive studies (Miller *et al.*,

TABLE I. Comparison of spectral metrics of vowels based on clustering by vowel category.

Metric	d'_{mn}	Rank	d'_{rms}	Rank	Average rank
$\log(F3/F2),\log(F2/F1)$	11.08	1	2.38	2	1.5
$\log(M3/M2),\log(M2/M1)$	9.86	2	2.50	1	1.5
$\log(K3/K2),\log(K2/K1)$	8.08	4	2.14	3	3.5
$(B3-B2),(B2-B1)$	8.58	3	1.73	5	4.0
$\log(B3/B2),\log(B2/B1)$	6.83	5	1.96	4	4.5
$(K3-K2),(K2-K1)$	6.67	6	1.28	6	6.0
$(M3-M2),(M2-M1)$	5.59	7	0.92	7	7.0
$(F3-F2),(F2-F1)$	2.87	8	0.25	8	8.0

1980, 1983), which indicated that logs of frequency ratios of the formants group the vowels as well or better than other metrics, and on the general importance of the log frequency scale in auditory science. With regard to the general prominence of frequency ratios, and thus the log frequency scale in auditory science, the following points are noted. Weber's law, which approximately describes frequency discrimination over a wide range of frequencies, implies a log frequency scale. In the case of music, it is true that all musical scales of all cultures are based on the octave and thus on log frequency scales. Chord structures are defined in terms of frequency ratios, and thus are most tractable in terms of log frequency. Additionally, transposition of melody is effective on the log frequency scale. Furthermore, Harris (1960) has shown that human listeners scale pitch in accordance with log frequency as long as the intervals to be scaled are small. A mel scale is only obtained when listeners are asked to scale large intervals. In the same article, Harris points out that the use of mel-like scales would result in very different musical forms based on relatively larger intervals than those commonly used, and, furthermore, such music would have to be monophonic, as mel-like scales will not allow harmony as it is presently conceived. From an anatomical point of view, it is noted that frequency-position maps of most mammals appear to be logarithmic over most of their range (Greenwood, 1961). Finally, the measurement of spectra in terms of log frequency provides the opportunity to compare distances and transition velocities in terms that can be easily compared to those of other sounds. All of the material cited above leads us to conclude that, even though competing scales may perform quite well in many circumstances as shown in the preceding material, the logarithmic frequency scale is quite fundamental to hearing, and that it is, therefore, the metric of choice until decisive results prove otherwise.

III. THE AUDITORY-PERCEPTUAL THEORY APPLIED TO VOWELS

The auditory-perceptual theory of phonetic recognition (Miller, 1984b, 1987a,b) will now be briefly described. It is comprised of a three-stage process (Miller, 1984a), where the acoustic waveform of speech is converted by the human listener to a string of category codes that correspond to the allophones of the language. The theory describes the "bottom-up" aspects of phonetic perception, and it is conceptual in nature. Stage 1 of the theory is the transformation of the acoustic waveform to auditory-sensory dimensions. Stage 2 is the transformation of the sensory data to perceptual values. In stage 3, the perceptual variables are converted to phonetic–linguistic categories.

In stage 1, it is assumed that short-term spectral analyses are performed on the incoming speech waveform. Each spectrum can be classified as a glottal-source spectrum, a burst-friction spectrum, or a combination of the two. At each moment, the spectral envelope patterns of the glottal-source and burst-friction sounds are represented as sensory responses or sensory pointers in a phonetically relevant auditory-perceptual space.[2] In stage 2, these sensory responses, or sensory pointers, are converted into a unitary perceptual response, or perceptual pointer, that is also located in the auditory-perceptual space. The perceptual response is a hypothetical construct or intervening variable that is based on the general notion that speech inputs are integrated to form a unitary perceptual stream. In practice, this hypothetical perceptual response is calculated by mathematically defined sensory-perceptual transformations that rely on the histories, trajectories, and dynamics of the sensory responses (Miller *et al.*, 1988). Finally, in stage 3, segmentation and categorization mechanisms that depend on the dynamics of the perceptual pointer in relation to perceptual target zones within the auditory-perceptual space result in a string of category codes that correspond to the allophones of the language. All of these concepts are elaborated as needed in the exposition that follows.

A. Auditory-perceptual theory applied to a sustained vowel

The analysis of a sustained vowel in terms of the auditory-perceptual theory will now be described. The vowel was intoned by a speaker, recorded, and digitally sampled at 20 kHz. It was then filtered with a high-pass filter set to 50 Hz to remove low-frequency noise. After high-frequency preemphasis, a short-term spectral analysis was performed. In our work, we routinely use a linear prediction analysis with 24 poles. A Hamming window of 24 ms was used and moved in 1-ms steps, and thus a spectral envelope was calculated for each millisecond of speech. In Fig. 10, an example of a spectral envelope from the vowel /IY/ is shown. In the auditory-perceptual theory, it is assumed that the listener's auditory system derives a sensory-spectral envelope that is nearly equivalent to the one shown. It is then assumed that this spectrum can be concisely represented as a sensory pointer in a phonetically relevant auditory-perceptual space (APS). The vowel spectrum is, of course, a glottal-source spectrum and will be represented as a glottal-source sensory pointer (GSSP) in the auditory-perceptual space. To locate the spectrum in the auditory-perceptual space, the center frequencies of the first three significant prominences are identified as sensory formant 1, $SF1$; sensory formant 2, $SF2$; and sensory formant 3, $SF3$. Furthermore, a sensory reference SR is located. As described in Appendix A, the sensory reference serves a variety of important functions in the auditory-perceptual theory, including talker normalization and the disambiguation of certain vowels. The sensory reference is believed to depend on the talker's average vocal

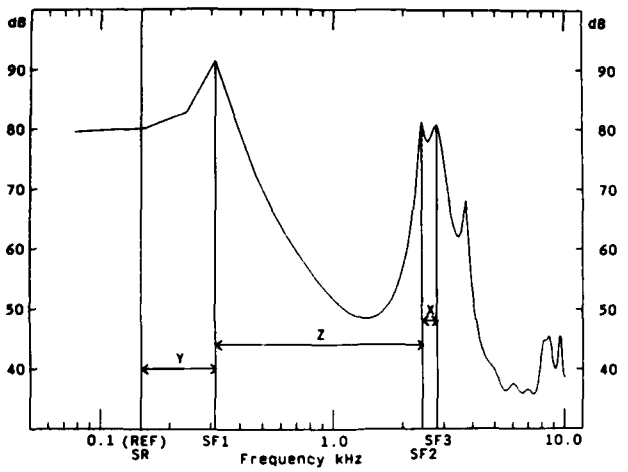


FIG. 10. A spectral envelope for the vowel /IY/ as intoned by a male speaker (JH). The sensory reference (*SR*) and the first three sensory formants (*SF*1, *SF*2, *SF*3) are located by the vertical lines. The values of the dimensions—*xyz*—of the auditory-perceptual space are indicated as the distances between the vertical lines. Since *GMF*0 = 120 Hz, *SR* = 150 Hz by Eq. (5). The values of the formants are *SF*1 = 310 Hz, *SF*2 = 2450 Hz, and *SF*3 = 2810 Hz. The corresponding values of *xyz* are 0.060, 0.315, and 0.898, respectively.

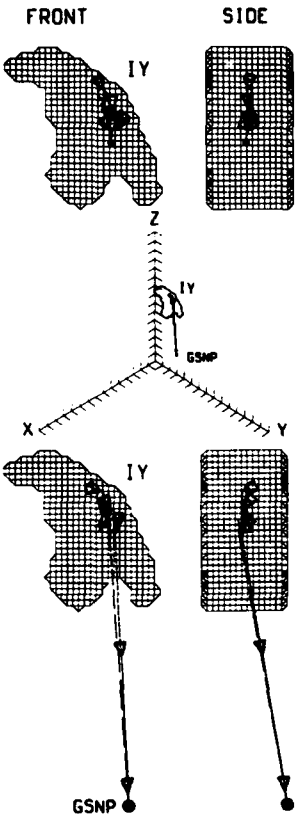


FIG. 11. Sensory (top) and perceptual (bottom) paths in the auditory-perceptual space are shown for the vowel /IY/ as intoned by a male speaker (JH). These are shown in front and side views. The envelope of the perceptual target zone for /IY/ is shown as is the location of the glottal source neutral point (GSNP). A pyramid is plotted on the path every 80 ms. The reduced figure (middle) shows the perceptual path with the locations of the relevant items in the APS. Here, *x*, *y*, and *z* axes are in 0.1 log units and the point of origin is (0, 0, 0).

characteristics and on appropriately filtered pitch modulations. For most of the purposes of the present paper, a simplified expression can be used. This is given by Eq. (5):

$$SR = 168(GMF0/168)^{1/3}, \quad (5)$$

where *GMF*0 is the geometric mean of the current speaker's voice pitch. For male speakers, a typical value of *GMF*0 is 132 Hz and the corresponding *SR* is 155 Hz. For female speakers, a typical value of *GMF*0 is 223 Hz and the corresponding *SR* is 185 Hz. For children 7–10 years of age, a typical value of *GMF*0 is 263 Hz and the corresponding *SR* is 195 Hz.

Returning to Fig. 10, we note that this speaker's *GMF*0 at the time of the spectrum was 120 Hz, and, therefore, the sensory reference was 150 Hz, as indicated in the figure. The formants *SF*1, *SF*2, and *SF*3 have center frequencies of 310, 2450, and 2810 Hz, respectively. This spectrum is located in the APS as follows. The variable *y* is defined as the logarithmic distance from *SR* to *SF*1:

$$y = (\log SF1) - (\log SR) = \log(SF1/SR). \quad (6)$$

The variable *z* is defined as the logarithmic distance from *SF*1 to *SF*2:

$$z = (\log SF2) - (\log SF1) = \log(SF2/SF1). \quad (7)$$

Finally, the variable *x* is defined as the logarithmic distance from *SF*2 to *SF*3:

$$x = (\log SF3) - (\log SF2) = \log(SF3/SF2). \quad (8)$$

The point associated with this spectrum has coordinates $x = 0.060$, $y = 0.315$, and $z = 0.898$. These coordinates define the position of the GSSP for this spectrum. Of course, during a sustained vowel, many such spectra are produced, and these are represented in the APS at the top of Fig. 11 in front and side views. Here, the sensory path of the GSSP is plotted for each millisecond of intoned vowel, and a directed pyramid is shown every 80 ms. The wire-mesh volume indicated in Fig. 11 is the perceptual target zone for the vowel /IY/. In our earliest work (Miller, 1984b), perceptual target zones were represented as spheres centered on points in APS corresponding to estimates of prototypical formant positions for each vowel. It was soon discovered that the data required the target zones to be large and irregularly shaped volumes. The particular zone shown for /IY/ in Fig. 11 is the result of our second attempt or second iteration in defining the perceptual target zones for simple vowels. (Such second-iteration zones are referred to as the I2 zones and are detailed in Sec. III C.) It is clear that this example falls inside of the I2 zone for the vowel /IY/.

In the auditory-perceptual theory, however, it is not the sensory path or its characteristics that directly determine the perceived vowel. Rather, it is the perceptual path that is hypothesized to be critical. A perceptual path always begins at the glottal-source neutral point, which represents an abstract home position for the perceptual response that is simply related to the locus of the spectrum associated with a uniform vocal tract. In the auditory-perceptual theory, every perceptual path begins and ends at the glottal-source neutral point (GSNP). The perceptual path is calculated from the sensory input. The glottal-source sensory pointer defined by the sensory formants and sensory reference drives the perceptual response or perceptual pointer (*PP*) through the APS in the manner of a second-order resonant system. It is as if the glottal-source sensory pointer and the perceptual pointer were joined by a spring and the APS offered viscous resistance. When the sensory pointer disappears, a similar, but much weaker, spring is activated that attracts the per-

ceptual pointer back to the glottal-source neutral point. The model assumes that sensory pointers and the glottal-source neutral point have very large masses relative to the perceptual pointer, which is arbitrarily assigned a mass of 1.0 unit. In this way, the perceptual response cannot act back on sensory inputs or the neutral point. The mathematical relations between the sensory pointer (sensory response) and the perceptual pointer (perceptual response) constitute an important part of the stage II or sensory-perceptual transformation of the auditory-perceptual theory (Miller *et al.*, 1988). Furthermore, the mathematics allows one to calculate values of the perceptual reference (*PR*) and the perceptual formants *PF* 1, *PF* 2, and *PF* 3. The perceptual path so generated in the APS is shown in front and side views at the bottom of Fig. 11. It is also shown in reduced size in the center of the figure to provide orientation with the axes of APS. It can be seen that the perceptual pointer leaves the glottal-source neutral point, enters the I2 zone for the vowel /IY/ where it remains for over 400 ms, and then returns to the GSNP.

In the auditory-perceptual theory, a perceptual target zone is activated to issue a neural symbol or category code corresponding to its associated allophone only when the perceptual pointer performs a segmentation maneuver within the zone. While the exact nature of the segmentation maneuver has yet to be formulated, it has been noted previously (Miller, 1984b; 1987a,b) that such maneuvers are associated with decelerations and relative lows in velocity of the perceptual pointer and with high curvature points along its path. In our sample case, the perceptual pointer decelerates within the target zone and exhibits a long period of low velocity. Also, the path shows high curvature as it goes to and returns from its average location inside of the perceptual target zone for /IY/. In terms of the auditory-perceptual theory, this stimulus would induce the perception of /IY/ because the perceptual pointer entered the perceptual target zone of /IY/ and executed a segmentation maneuver that caused the target zone to be activated. Also, it is assumed that the activation of the target zone is maintained during the sustained low-velocity part of the path.

It is also necessary to model the loudness of the perceptual response, as well as its spectral characteristics. For this purpose, it is assumed that the loudness of the perceptual pointer rapidly grows over some 10–70 ms at the beginning of a utterance and decays more slowly, perhaps over a period of 100–250 ms, after the cessation of the sensory input (Miller *et al.*, 1988). The values for the growth and decay are consistent with knowledge of loudness integration (Sharf, 1978) and forward masking (Zwislocki, 1978). In the simple case of the sustained vowel illustrated in Fig. 11, the implication is that the loudness of the perceptual pointer decays to zero as the perceptual pointer returns to its home position, the glottal-source neutral point. More generally, it is hoped that the growth and decay of loudness may help to explain why speech can be "heard" during brief silences.

B. Auditory-perceptual theory applied to a consonant-vowel syllable

A slightly more complicated and interesting example of the application of the auditory-perceptual theory will now be described. In this case, a female speaker uttered the syllable /LAA/. The sensory path is shown in APS at the top of Fig. 12 in front and side views. The I2 zones for the consonant /L/ and the vowel /AA/ are shown. The resultant perceptual path is shown at the bottom of the same figure. As can be seen, the perceptual pointer follows a path from the GSNP to the /L/-target zone, to the /AA/-target zone, and, then, back to the GSNP. Segmentation maneuvers, such as decelerations, periods of low velocity, and sharp turns, are assumed to activate first the /L/-target zone and then the /AA/-target zone.

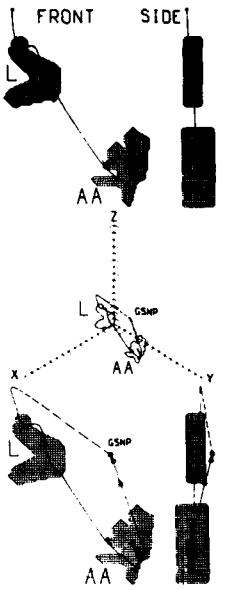


FIG. 12. Sensory (top) and perceptual (bottom) paths in the auditory-perceptual space for the syllable /LAA/ produced by a female speaker (LT). These are shown in front and side views. There is a break in each path at each millisecond. A pyramid is plotted on the perceptual path every 80 ms. Preliminary estimates of the target zones for the /L/ (light l) and the vowel /AA/ are shown. The perceptual path can be interpreted as entering and activating these zones. The reduced (middle) figure shows the perceptual path with the locations of the relevant items in the APS.

C. Vowels in the APS

Since the patterns of spectral change associated with vowels are generally rather slow, it is often the case that the perceptual and sensory paths will be identical, or nearly so, in those segments of the paths that trigger the perception of the vowel. Therefore, it is reasonable to estimate the positions of the perceptual pointer associated with the vowels directly from formant values reported in the literature, in which case the judgments of the authors as to whether the reported formant values are representative of a vowel are accepted. Where the pitch is not given, values of 133 Hz for adult males, 225 Hz for adult females, and 263 Hz for children 7–10 years of age are assumed. In our own work, we listen to windowed segments of a path, examine patterns of velocity and curvature in the APS, and finally average points over a segmented region of the perceptual path. In this way, 406 data points were collected for the simple, nonretroflex vowels of American English: /IY/—24, /IH/—88, /EH/—89, /AE/—23, /AA/—24, /AO/—25, /AH/—87, /UH/—24, and /UW/—22. These include data for men, women, and children in a variety of phonetic contexts and speaking rates and from a variety of laboratories. All are from single syllable utterances and the sources of these data are described in detail in Appendix B.

It was immediately noticed that the nonretroflex vowels fall in a narrow slab in the APS. This is illustrated by show-

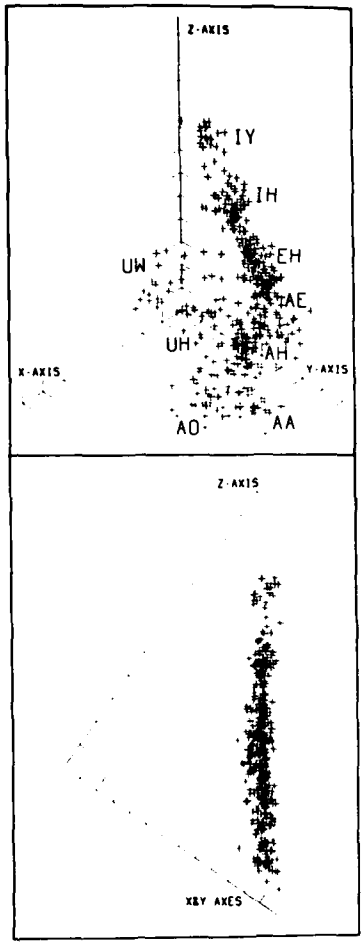


FIG. 13. The 406 examples of the nonretroflex vowels used to develop the I2 target zones are shown in front (top) and side (bottom) views. The regions associated with vowels are indicated by the labels in the front view. The side view dramatically illustrates the existence of the vowel slab. For the side view, the axes of the APS have been rotated to bring the vowel slab to the vertical. In this particular orientation, the x and y axes appear to superpose. For both panels, x, y, and z axes have tic marks every 0.1 log units and the point of origin is $(0, 0, 0)$.

ing all 406 points in front (top panel) and side (bottom panel) views in Fig. 13. The middle of the vowel slab is defined by the sum plane where $x + y + z = 1.220 \pm 0.135$. This means that the slab has a thickness[3] of about 0.156 log units or about one-half octave. The existence of the vowel slab is almost certainly related to vocal tract acoustics as expressed in relations between vocal tract size, fundamental frequency, and the value of the third formant. A related, but distinct, geometric approach to vowel formant frequencies can be found in Broad and Wakita (1977).

Sometimes it is useful to rotate the axes so that the vowel slab is brought to the vertical. This can be accomplished by the equations given below:

$$x' = 0.7071(y - x), \tag{9}$$

$$y' = 0.8162z - 0.4081(x + y), \tag{10}$$

$$z' = 0.5772(x + y + z). \tag{11}$$

Generally, we speak of space APS as having coordinates xyz and space SLAB as having coordinates $x'y'z'$. Notice that this transformation does not translate the origin but only rotates the axes. Space SLAB is often convenient for examining vowels, while space APS is often more convenient for examining other allophones. Many times the xyz axes are shown, but the display is rotated so that the vowel slab is in the vertical.

After collecting the data points shown in Fig. 13 and described in Appendix B, we developed target zones, the I2 zones, to represent the clustering of the vowels into distinct regions of the APS. Since most of the variability of the points was along the height (y' axis) and width (x' axis) of the vowel slab and not in the depth (z' axis), the target zones were constructed with variable outlines in the height and width dimensions, but with uniform, that is, rectangular, outlines in the depth dimension. The author used a computer-aided method with a resolution of 0.01 log units to draw these zones by hand. He also imposed additional constraints on the regularity of $x'y'$ borders, since the density of the data points was not sufficient to constrain the outlines. Also, he ignored certain points because they seemed to be unlikely and were thought to be errors. In spite of the fact that certain points were excluded in creating the zones, *all* data were retained for purposes of display and evaluation. By this process, a set of nonoverlapping perceptual target zones was created that classified the data points with good accuracy.

The outlines of hand-drawn target zones are shown in Fig. 14. Here, the slab coordinates y' and x' [see Eqs. (10) and (9)] are shown, and the triangular outline is the intersection of the slab coordinates with the original xyz coordinates at the middle of the vowel slab. The maplike character of vowel regions is apparent. The hand-drawn zones were given depth limits in z' that were consistent with the data and then were subjected to a software contouring process that created wire-mesh enclosures that fit the hand-drawn contours with an error of ± 0.01 log units. These wire-mesh

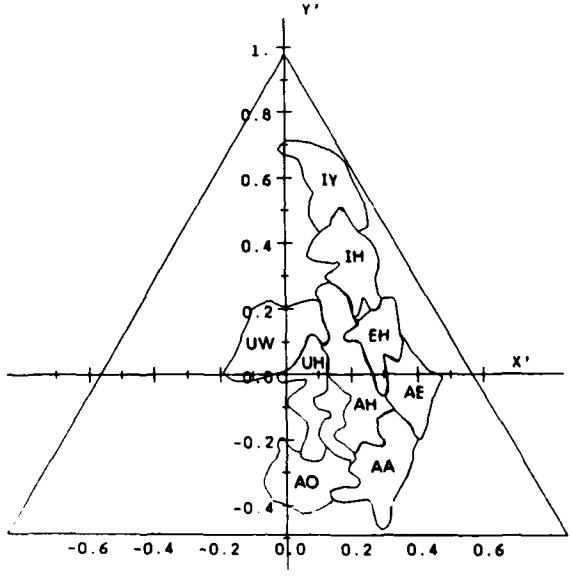


FIG. 14. The second iteration (I2) target zones for the nine nonretroflex vowels of American English are shown in SLAB coordinates [see Eqs. (9)–(11)]. In this orientation, the z' axis is perpendicular to the $x'y'$ plane. Note the maplike character of these zones with their large areas, irregular shapes, and abutting boundaries.

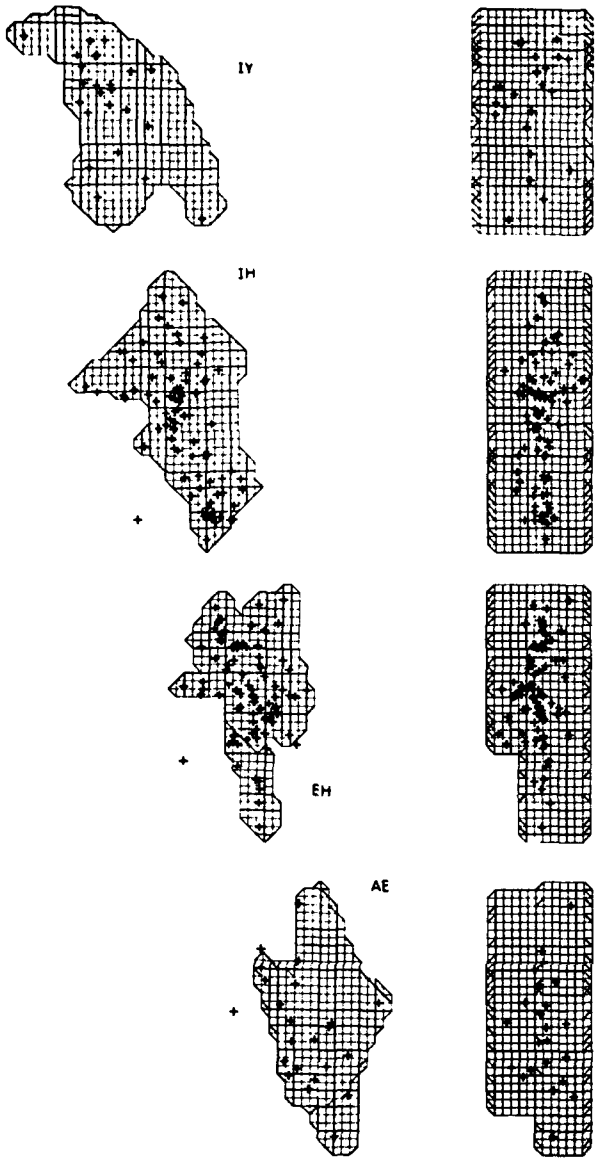


FIG. 15. Data points and target zones for the vowels /IY/, /IH/, /EH/, and /AE/ are shown in front (left) and side (right) views. The numbers and percentages of points enclosed in each wire-mesh target zone can be visualized and compared with the numbers presented in Table II. See Fig. 14 for the location of the individual target zones in APS.

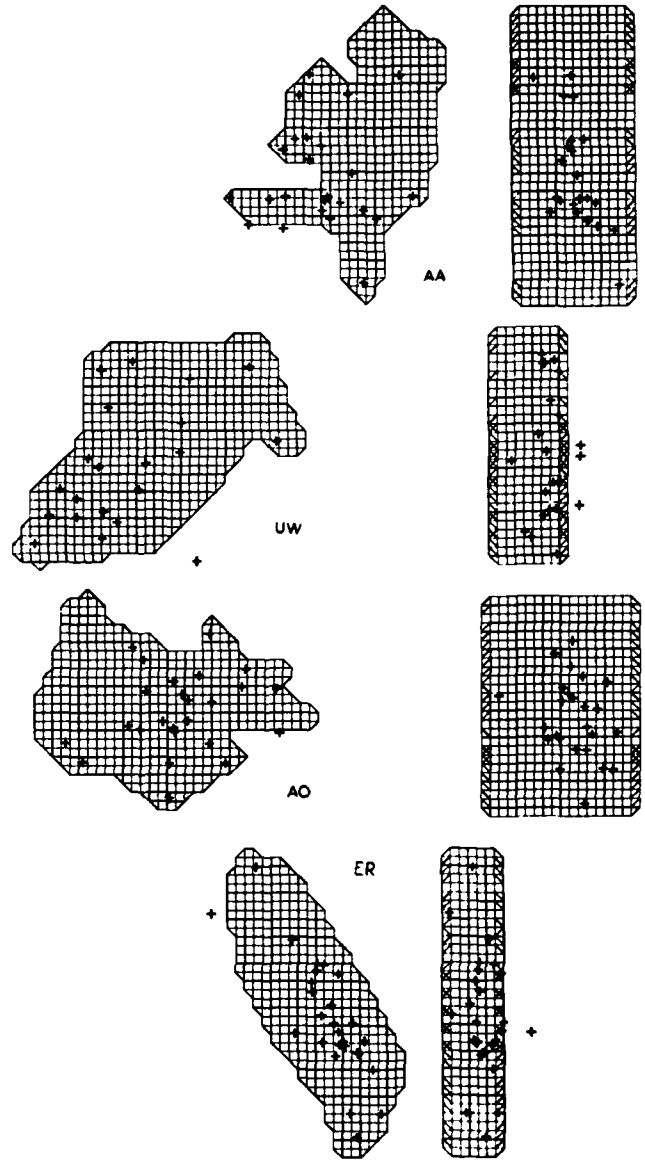


FIG. 16. Data points and target zones for the vowels /AA/, /UW/, /AO/, and /ER/ are shown in front (left) and side (right) views. The numbers and percentages of points enclosed in each wire-mesh target zone can be visualized and compared with the numbers presented in Table II. See Fig. 14 for the location of the individual target zones in APS.

enclosures are shown in front view ($x'y'$) and side view ($y'z'$) in Figs. 15–17. Note the irregular borders in the front view and the rectangular borders in the side view.

A sense of the accuracy with which these preliminary I2 target zones classify the vowel data can be obtained from Table II and from Figs. 15–17. Here, all of the 406 data points for the oral, nonretroflex vowels are shown, as well as an additional 29 samples of the vowel /ER/ for a total of 435 points. Overall, the present target zones segregated the vowels of American English with 93% accuracy. Correct classification ranges from 100% for /IY/ to 82% for /UW/. Table II demonstrates that a high proportion of the points falls in the appropriate target zone or in adjacent unclaimed regions that could be easily incorporated into the target zone. These data are in accord with a variety of studies showing that human listeners identify English vowels in CVC context with approximately 95% accuracy (see, for example, Peterson and Barney, 1952; Gottfried and Strange, 1980; Macchi, 1980). Inspection of Figs. 15–17, additionally, shows that most of the errors or misses are "near misses," which can be judged by the fact that the wire-mesh enclosures illustrated in the figures consist of squares that are 0.01 log units on a side, or approximately 1/30 of an octave. About one-half of the errors fall between 0 and 0.03 log units, that is, fall within 1/10 of an octave of a target zone boundary.

These zones are characterized as being very large with irregular and abutting boundaries. The boundary regions seem to be extremely small in relation to the size of the zones. It is the hypothesis of the auditory-perceptual theory that target zones with the qualities just described will accurately describe the simple vowels of each dialect of a spoken language. However, in order to accurately define these zones,

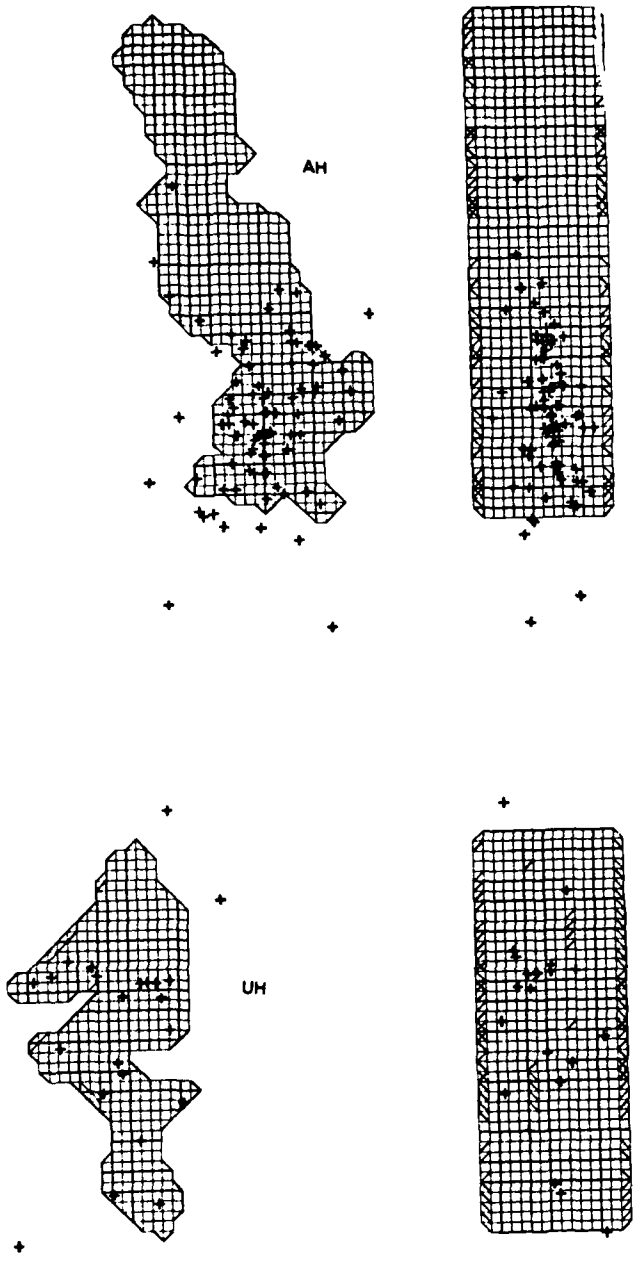


FIG. 17. Data points and target zones for the vowels /AH/ and /UH/ are shown in front (left) and side (right) views. The numbers and percentages of points enclosed in each wire-mesh target zone can be visualized and compared with the numbers presented in Table II. See Fig. 14 for the location of the individual target zones in APS.

we will have to make improvements in our procedures and database. These include: (1) better procedures for the segmentation of perceptual paths, (2) better procedures for listener verification of points in unbiased listening tasks, (3) procedures for contouring sets of like-identified points that are more precise and objective than present techniques, and (4) much more data. Since the 12 zones were created, we have been making progress toward these required improvements (Miller, 1987c).

The large sizes of the target zones, along with their irregular boundaries, may prove useful in accounting for much of the measured variability in acoustic spectra associated with vowels. We are currently conducting experiments using vowels synthesized from *xyz* values from APS to verify the perceptual significance of the present irregularly shaped target zones. It is clear that each vowel is represented by a very large class of spectral patterns. The auditory-perceptual equivalent of Lloyd's claim that like formant ratios produce like vowel percepts is that like values of *xyz* produce like vowel percepts. This, of course, implies that like values of the ratios ($SF1/SR$), ($SF2/SF1$), and ($SF3/SF2$) produce like vowel percepts. But one must not make the logical error of inferring that like percepts have like values of *xyz*, since the large sizes of the perceptual target zones disprove such an inference even as an approximation. It is a hypothesis of the auditory-perceptual theory that the large and irregular shapes of the perceptual target zones will account for a substantial proportion of differences in vowels due to coarticulation effects as well as a substantial proportion of the variability due to differences between talkers in articulatory style and habit. It also appears that the sensory reference concept helps to eliminate differences among talkers due to size, age, and gender, as well as serving to disambiguate the vowels with similar $F2/F1$ and $F3/F2$ ratios noted in Sec. I. Therefore, the auditory-perceptual approach may lead to a relatively simple account of talker differences and coarticulation effects in vowel perception.

IV. OTHER DESCRIPTIONS OF THE VOWELS AS RELATED TO THEIR LOCI IN THE AUDITORY-PERCEPTUAL SPACE

The vowels have often been described in terms of dimensions of articulation. For example, charts in *The Principles of the International Phonetics Association* (International Phonetics Association, 1949) show the loci of vowels in relation to a high–low dimension and a front–back dimension, which represent the position of the "tongue hump." In an alternate presentation, the front–back dimension is plotted against an open–close dimension that represents the jaw opening associated with the vowel. These charts are based in a large part on the x-ray photographs of Jones (Jones, 1914, 1919). Another description of the vowels in terms of variables related to acoustic timbre has been suggested by Fant (1973). In this scheme, vowels are classified along a grave–acute dimension and a compact–diffuse dimension. Graveness is related to the average position of $F1$ and $F2$ prime on a mel scale, while compactness is related to separation between $F1$ and $F2$ prime similarly expressed in mels (Fant, 1973, p. 188). When the oral, nonretroflex vowels of American English are viewed in the vowel slab, relationships similar to those seen in the planes just described are observed. In fact, we have noticed that very simple rotations of the vowel slab produce relationships similar to those of either the articulatory or timbre dimensions. These similarities are illustrated in the four panels of Fig. 18. For these figures, the average values of *xyz* were calculated for each vowel from the mean data for men, women, and children of Peterson and Barney (1952). The several organizations of the nonretroflex vowels of American English shown in Fig. 18 are simple rotations of the original *xyz* dimensions of APS. To aid in illustrating this point, the intersections of the coordinate planes

TABLE II. Number of tokens of each vowel enclosed by each target zone. Here, L is the target zone for light-l, AUC is the adjacent unclaimed region, NAUC is the nonadjacent unclaimed region, and the numbers in parentheses include points in AUCs as correct.

Vowels	Target zones													Total	Proportion correctly enclosed	
	IY	IH	EH	AE	AA	AO	AH	UH	UW	ER	L	AUC	NAUC			
IY	24	...	...	...	...	...	...	...	...	...	...	...	...	24	1.00	
IH	...	87	...	...	...	...	1	...	...	...	...	...	...	88	0.99	
EH	...	...	87	...	...	...	1	...	...	...	...	1	...	89	0.98	(0.99)
AE	...	...	2	21	...	...	...	...	...	...	...	...	...	23	0.91	
AA	...	...	...	...	22	...	...	...	...	...	...	2	...	24	0.92	(1.00)
AO	...	...	...	...	1	24	...	...	...	...	...	...	...	25	0.96	
AH	...	...	1	1	3	1	73	2	...	...	...	6	...	87	0.84	(0.91)
UH	...	...	...	...	...	1	2	20	...	...	...	...	1	24	0.83	(1.00)
UW	...	...	...	...	...	...	...	1	18	...	3	...	...	22	0.82	
ER	...	...	1	...	...	...	...	...	...	27	...	1	...	29	0.93	(0.97)
Total	24	87	91	22	26	26	77	23	18	27	3	10	1	435	92.6	(94.9)
														Total Correct	403	(413)

of APS, that is, the xy, xz, and yz planes, with the illustrated plane are shown by the triangle plotted on each panel. The upper left panel shows the vowel plane in slab coordinates. In this case, $x'y'z'$ are as defined in Eqs. (9), (10), and (11), where z' is fixed at 0.704.

The front–back and high–low dimensions of panel (b) can be calculated as

$$B = -0.3536(y-x) - 0.7068(z) + 0.3534(x+y), \quad (12)$$

$$H1 = 0.4081(z) - 0.2041(x+y) - 0.6123(y-x), \quad (13)$$

where B is back and $H1$ is high in this rotation.

The dimensions shown in panel (c) are given by

$$O = 0.3536(y-x) - 0.7068(z) + 0.3534(x+y), \quad (14)$$

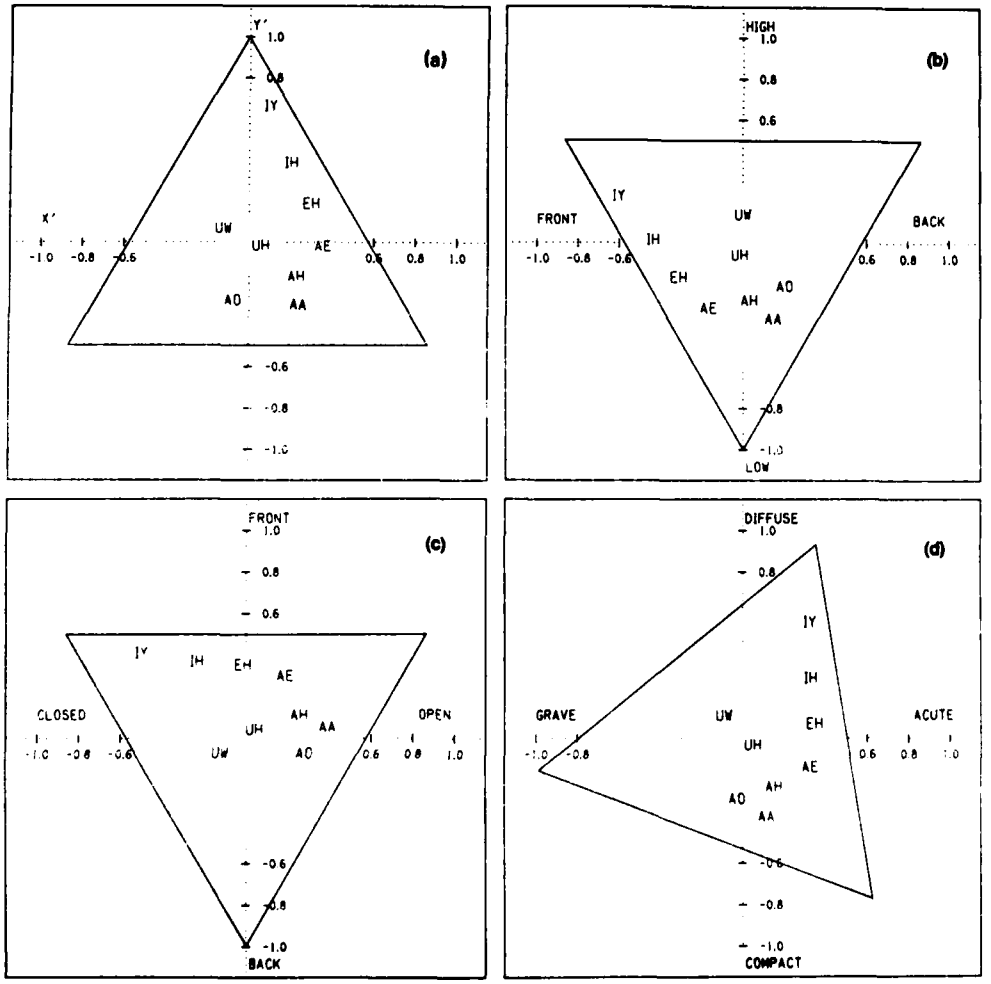


FIG. 18. The locations of the oral, nonretroflex vowels of American English are shown in four coordinate systems that are simply rotations of the auditory-perceptual space. In panel (a), the SLAB coordinates defined by Eqs. (9)–(11) are used. In panel (b), the articulatory dimensions of Eqs. (12) and (13) are used. In panel (c), the articulatory dimensions of Eqs. (14) and (15) are used. Finally, in panel (d), a rotation [Eqs. (16) and (17)] is used that provides dimensions similar to the grave–acute and compact–diffuse dimensions of Fant (1973). The symbols are located at the mean positions of the combined data for men, women, and children of Peterson and Barney (1952).

TABLE III. Transformations of xyz to other dimensions for description of the vowels.

Variable[a]	x	y	z
x'	−0.7071	0.7071	0.0000
y'	−0.4081	−0.4081	0.8162
B	0.7070	−0.0002	−0.7068
$H1$	0.4082	−0.8164	0.4081
O	−0.0002	0.7070	−0.7068
$H2$	−0.8164	0.4082	0.4081
A	−0.8065	0.5139	0.2925
D	−0.1273	−0.6344	0.7620

[a] See text and Eqs. (9)–(17) for definitions.

$$H2 = 0.4081(z) - 0.2041(x+y) + 0.6123(y-x), \quad (15)$$

where O is open and $H2$ is high in this context.

Finally, in the lower right panel [panel (d)], horizontal and vertical axes are

$$A = 0.6602(y-x) + 0.2925(z) - 0.1463(x+y), \quad (16)$$

$$D = 0.7620(z) - 0.3810(x+y) - 0.2534(y-x), \quad (17)$$

where A is acute and D is diffuse.

These same transformations are given in another form in Table III. Each variable is calculated by multiplying the values of xyz by the indicated coefficients and summing the resulting products.

Other articulatory descriptions of vowels can be related to the dimensions of the APS. Consider the retroflexion of vowels. As is well known, $SF3$ drops close to $SF2$ for the /ER/ vowel of American English, and both approach a value of about 1510 Hz for males, 1793 Hz for females, and 1983 Hz for children (Peterson and Barney, 1952). In terms of the dimensions of APS, this means that the values of x and the value of the sum $(x+y+z)$ will decrease during retroflexion of a vowel. This further implies that the /ER/ vowels will fall behind the vowel slab defined for the nonretroflex vowels. This is the case. We have used 29 values of /ER/ in creating the I2 zone for /ER/ (see bottom of Fig. 16). As illustrated in Fig. 19, these points fall behind the vowel slab. The average value of $(x+y+z)$ was 1.22 for the nonretroflex vowels, while the average for /ER/ was 1.02. Only a few tokens of the vowels /UW/ and /AO/ have values of $(x+y+z)$ as small as the largest values for tokens of /ER/. However, in regions of the vowel slab close to the /ER/ region, all oral vowels have values of $(x+y+z)$ that are larger than that of any /ER/ token. Thus, it is clear that the retroflex vowel /ER/ falls behind the nonretroflex vowels of American English.

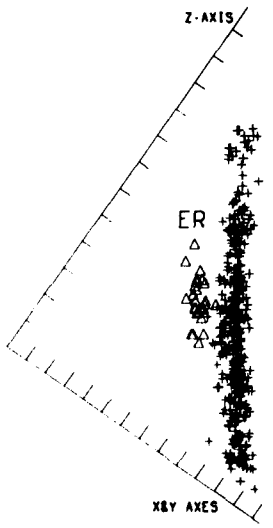


FIG. 19. Side view of the vowel slab showing that most of the 29 data points for the retroflex vowel /ER/ (triangles) fall behind the data points for the oral, nonretroflex vowels (crosses). In this orientation, the x and y axes appear to superpose. Here, x, y, and z axes have tic marks in 0.1 log units and the point of origin is (0, 0, 0).

Furthermore, it is possible to calculate retroflexion as a function of the variables SR, $SF2$, and $SF3$. If one defines K as constant with a value of about 7.8, retroflexion should peak as the geometric mean of $SF2$ and $SF3$ approaches KSR and as $SF2$ and $SF3$ approach equality. Otherwise said, as $SF2$ and $SF3$ approach 1510, 1793, and 1983 Hz for men, women, and children, respectively, retroflexion is indicated.

The measure of nasalization utilized in the auditory-perceptual theory is based on the observations of Fant (1970), Fujimura (1962), and Stevens *et al.* (1987). They show that a complex interaction of the nasal pole and nasal zero with the first resonance of the oral tract results in two peaks in the lower part of the spectrum. While, for oral vowels, these two peaks are merged, during nasalization we define the lower of these two peaks as the low first sensory formant ($SF1L$) and the higher of these two peaks as the high first sensory formant ($SF1H$), independent of their bases in vocal tract acoustics. Such interpretation is simple in the case of nasalization of a vowel with a high $F1$. Here, the nasal pole corresponds to $SF1L$ and the nasalized $F1$ corresponds to $SF1H$, while the nasal zero is associated with the valley between the two and shifts the nasalized $F1$ up from its oral position. The case of the nasalization of a vowel with a low $F1$ is different. Here, the nasal pole and oral $F1$ can interact to produce a peak corresponding to $SF1L$. The nasal zero interacts with the falling spectrum to produce, with increasing frequency, a valley and then a local maximum in the spectrum. This local maximum is taken as $SF1H$, even though it may have no corresponding resonance in the vocal tract. Otherwise said, $SF1L$ and $SF1H$ are defined as the two low-frequency peaks that appear in the spectral envelopes of nasalized sonorants.

The degree of nasalization then is described by the equation

$$N = \log(SF1H/SF1L), \quad (18)$$

where $SF1L$ is the low first formant and $SF1H$ is the high first formant. In oral vowels, N has an effective value of zero. When nasalization is present, the dimensions of the APS are given by $x = \log(SF3/SF2)$, $y = \log(SF1L/SR)$, and $z = \log(SF2/SF1H)$. By this formulation, $(x+y+z)$ becomes small during nasalization because, in essence, the value of N is subtracted from the sum. That is, in the oral case, $(x+y+z) = \log(SF3/SR)$, while, in the case of nasalization,

$$x+y+z = \log\frac{SF3}{SR} - \log\frac{SF1H}{SF1L} = \log\frac{SF3}{SR} - N. \quad (19)$$

Thus nasalized vowels will fall well behind the vowel slab when N is large.

Finally, consider lip rounding. Fant (1970) has noted that lip rounding tends to lower the formant pattern. We

have observed in comparing /UW/ with /IY/ in American English that the rounded /UW/ tends to be toward the back portion of the vowel slab, while the spread vowel /IY/ tends to be toward the front portion of the vowel slab. Jongman and Fourakis (1987) have noted a similar trend in comparing rounded and unrounded vowels in German. Thus rounding is indicated by a movement *within* the vowel slab from larger values of $(x+y+z)$ to smaller values of this variable. A pure measure of rounding in terms of the auditory-perceptual variables that is uncontaminated by other variables, such as nasalization and retroflexion, has not yet been derived, but it seems likely that such a measure can be found.

In this section, we have indicated how the variables of the auditory-perceptual theory can be related to other, perhaps more familiar, descriptors of the vowels. Consideration of these relations suggests that five variables are necessary to disambiguate the vowels. These are SR, $SF1L$, $SF1H$, $SF2$, and $SF3$. With these descriptors of the acoustic spectrum, it appears that one can characterize vowels with regard to retroflexion, nasalization, and rounding, as well as with regard to tongue position and jaw opening. Furthermore, the five variables described above can be combined into the three dimensions of the APS to provide a concise description of vowel spectra that can be easily visualized.

V. EXTENSIONS OF THE AUDITORY-PERCEPTUAL APPROACH TO OTHER ISSUES IN VOWEL PERCEPTION

In this section, extensions of the auditory-perceptual approach to other issues in vowel perception are briefly indicated. Specifically, issues of the relationship of voice pitch to perceived vowel category, of the transposition experiment, of diphthongs and diphthongization of vowels, and of the nature of the boundaries and vowel target zones are discussed.

A. Voice pitch and perceived vowels

It is well known that, under most conditions, the identity of a perceived vowel depends strongly on the formant values of the spectrum and is independent of voice pitch. However, under certain circumstances the voice pitch can influence a vowel's identity, even when the formants are fixed (Fant *et al.*, 1974; Miller, 1953; Fujisaki and Kawashima, 1968; and Scott, 1976). Within the auditory-perceptual theory, two hypothetical mechanisms are included that would allow voice pitch, under certain circumstances, to control the perception of vowel category. One mechanism follows from the fact that a change in the geometric mean of the fundamental ($GMF0$) results in a change in the sensory reference (SR)—Eq. (5)—and thus the value of y, since $y = \log(SF1/SR)$. Therefore, a change in $GMF0$ can cause a spectral representation to shift its location in the APS. Since the target zones are large and changes in $F0$ are compressed by Eq. (5), most changes in $GMF0$ will not change vowel identity, and this is concordant with everyday experience. If, however, the spectrum lies near a boundary of a vowel target zone, then an appropriate shift in $GMF0$ can cause the locus of the spectrum in APS to shift from one target zone to another, thus changing the identity of the perceived vowel.

A second mechanism proposed in theory is to have the SR influenced by pitch modulations. This hypothesis is motivated by the well-known relationships of pitch modulation to stress and a desire, then, to incorporate such modulations in the sensory and perceptual paths of an utterance. In this version of the theory,

$$SR = 168(GMF0/168)^{1/3} + \mathrm{FIL}(\mathrm{mod}\ F0), \qquad (20)$$

where the term $\mathrm{FIL}(\mathrm{mod}\ F0)$ is meant to represent the modulations in $F0$ after bandpass filtering. By this approach, the SR would not be influenced by the very slow changes in $F0$ associated with pitch declination (Cooper and Sorensen, 1981), nor would it be influenced by the rapid changes that occur at voice onset and offset. However, appropriate pitch modulations could shift the location of a spectrum in APS, and if such a shift crossed a vowel boundary, then the identity of the perceived vowel would be changed. Quantitative evaluation of this hypothesized relation is a matter for future attention.

B. Transposition of vowel spectra

Studies of the effects of transposition of vowel spectra began with mechanically synthesized vowels in the 1800s, and their number was greatly increased as Edison's phonograph became available in laboratories. Among many publications on, or related to, this topic are those of Bell (1879), Lloyd (1890a,b, 1891, 1892, 1894), Pipping (1893), Fletcher (1929), Chiba and Kajiyama (1941), Ochai *et al.* (1955), Kurtzrock (1956), Tiffany and Bennett (1961), Klumpp and Webster (1961), Slawson (1968), Daniloff *et al.* (1968), Fujisaki and Kawashima (1968), and Sekimoto (1982). Now, if the formant-ratio theory of vowel perception were strictly true, vowel identity would be maintained as the vowel spectrum is transposed up or down the logarithmic frequency axis. The experimental results cited above confirm this prediction within limits, and the remaining issues relate to the range of transposition and the causes of its limits. Early apparatus, such as the Ediphone used by Bell (1879), suffered serious distortion as the range of transposition was increased. Later work is less subject to interpretation in terms of distortions produced by the experimental procedures. One reasonable hypothesis is that transposition is limited by reduced resolving power of the auditory system in the low and high ranges, as is exemplified in the increase in the Weber fraction for frequency discrimination below 500 Hz and above 2000 Hz, as measured by Wier *et al.* (1977). Such reductions in resolving power may cause certain spectral peaks to fuse and either shift the identification or cause the vowel to become indistinct. This view is partially supported by the finding that women's speech can tolerate more downward transposition than men's speech, while men's speech can tolerate more upward transposition than women's (Klump and Webster, 1961). Another possible limitation on the transposition experiment follows from the auditory-perceptual theory. In order to maintain vowel identity in the auditory-perceptual theory, the ratio of $F1$ to SR must be maintained such that a vowel boundary is not crossed. Since most transposition experiments have not made changes in $F0$ relative to $F1$ so as to maintain appropriate

$F1/SR$ ratios, this condition has not been met. Therefore, some of the observed failures of transposition may relate to changes in the $F1/SR$ ratio. Perhaps future research will clarify this possibility.

Even though all of the conditions for successful transposition are not yet known and further experimental work can be of interest, the results of the cited experiments in and after the 1950s all demonstrate that transposition holds reasonably well over the ranges normally encountered in the vowel sounds of children, women, and men. Such results are consistent with the auditory-perceptual interpretation of the vowel.

C. Vowels with changing formants

Naturally produced vowels often include formant movements as discussed by Nearey and Assmann (1986), Nearey (1989), and Strange (1989). Of course, it has been well known that such movements play a role in the perception of diphthongs and in the perception of obviously diphthongized simple vowels. In the recent work of Nearey and Strange, however, an emphasis has been placed on the possible importance of such movements in the perception of the simple vowels even when they are perceived as such. In the auditory-perceptual theory, diphthongs or diphthonglike sequences are to be treated either as phone pairs or phone-plus-glide sequences, as discussed elsewhere (Miller 1987a,b; Chang, 1987b), and in these cases it is clear that formant movements play an important role in controlling which phonetic elements are perceived. But, it is also true in the case of simple vowels that the auditory-perceptual theory specifies that the *path* of the perceptual pointer and the *dynamics* of that pointer in relation to the perceptual target zones control the perceived category of the vowel. While all the implications of the theory in relation to these factors have not been explored, such work is under way in our laboratory.

D. Boundaries of vowel target zones

The boundaries of the perceptual target zones for the vowels, as illustrated in the figures of this paper, appear to be fixed and rigid. In the theory of bottom-up processing of speech provided by the auditory-perceptual theory, this is a correct interpretation. Indeed, the boundaries of the perceptual target zones of adults are assumed to be stable. We are aware that this view of the boundaries of vowels appears to contradict many published observations, for it is clear that estimates of vowel boundaries determined with synthetic stimuli are not fixed but can be shifted in location under the influence of a variety of variables. These include: (1) stimulus range (Ades, 1977; Goldberg, 1986), (2) adaptation and anchor effects (Morse *et al.*, 1976; Sawusch and Nusbaum, 1979; Repp *et al.*, 1979; Sawusch *et al.*, 1980; Fox, 1985), (3) order effects (Fry *et al.*, 1962; Repp *et al.*, 1979; Cowan and Morse, 1986), (4) contextual effects (Millar and Ainsworth, 1972; Repp *et al.*, 1979; Macchi, 1980; Gottfried *et al.*, 1985), and (5) experimental design and methods employed (Macmillan *et al.*, 1977, 1987; Ades, 1977; Pisoni, 1973). Even though a number of these studies provide conflicting results on similar issues, the overall implication that the apparent boundaries of vowel categories can be shifted by a variety of experimental factors is conclusively demonstrated. Therefore, avenues for the explanation of changes in boundaries must be provided.

One such avenue is through "top-down" processing. As suggested elsewhere (Miller, 1987b), top-down processing can be easily integrated into the auditory-perceptual theory in a variety of ways. One such mechanism provides for top-down processing driven by situational contexts, such as task demands and expectations, to influence the path of the perceptual pointer. By this mechanism, the boundaries of the perceptual target zones remain fixed, and it is the path of the perceptual pointer that changes. In this way, it is only the boundary "as calculated" that changes, while the "true" inherent boundary remains fixed. In another explanation of observed boundary shifts, it is proposed that temporary target zones can be developed based on the stimuli being presented during an experimental session (Miller, 1987b). Of course, whether such explanations prove useful is a matter for future research. What is of current interest is that the auditory-perceptual theory posits that boundaries of target zones for the vowels of an adult listener's native language are fixed, and that these are the boundaries that are relevant to the bottom-up processing of clear speech in everyday listening situations. On the other hand, under conditions other than those just described, the apparent boundaries of the perceptual target zones can be shifted by a variety of factors, as conclusively demonstrated in the cited literature.

VI. CONCLUDING COMMENTS

The auditory-perceptual interpretation of vowel perception has been described. Within this framework, the sensory and perceptual paths provide a detailed description of the changes in spectral patterns of the formants during the course of an utterance. It is proposed that a segmentation mechanism based on the dynamics of these spectral patterns causes perceptual target zones to issue neural symbols or category codes that correspond to the vowel sounds. Furthermore, it is demonstrated that the preliminary target zones described in this paper can classify the present database of American English vowels with 93% accuracy.

The auditory-perceptual theory is a descendent of formant-ratio theory and attempts to deal with the latter's shortcomings. A sensory reference concept is introduced that not only serves to improve talker normalization, but also makes the interpretation of the $F2/F1$ and $F3/F2$ ratios contingent on the $F1/SR$ ratio. Since SR is a relatively weak function of the average $F0$ (pitch), this concept brings a factor akin to the absolute frequency location of the formants into the theory as a determiner of vowel quality, and this appears to allow the disambiguation of vowels that have similar $F2/F1$ and $F3/F2$ ratios. The auditory-perceptual theory also offers at least a partial solution to the coarticulation problem by positing very large target zones with irregular borders. In this way, very large differences can exist among vowel spectra that are all associated with the same vowel phoneme. It is also proposed that careful study of the dynamics of spectral change will reveal the hypothesized segmentation maneuvers that activate the perceptual target

zones. Furthermore, a mechanism where voice pitch can alter the perceived vowel category is described. Finally, the relations between articulatory descriptions of the vowel, including tongue body locations, retroflexion, nasalization, and lip rounding with the loci of the vowels in the auditory-perceptual space, are discussed and shown to be amenable to quantitative description in the APS.

Even though further evaluation and refinement of the concepts introduced in this paper are needed before the merit of the auditory-perceptual approach can be decisively assessed, it does seem to provide a useful organizing framework for further study of the acoustic and auditory correlates of vowel sounds.

ACKNOWLEDGMENTS

This work was supported by NIH Program Project Grant NS 03856, NIH Grant NS 21994, and AFOSR Grant 86-0335 to Central Institute for the Deaf. Many people have worked with me at one time or another over the past several years on the topic of vowel perception and analysis. Included in this group are: A. Maynard Engebretson, N. Rao Vemula, Deborah G. Servi, Nancy Dunlop, Robert H. Gilkey, and Arnold H. Heidbreder. More recently, Marios S. Fourakis, Allard Jongman, Steven J. Sadoff, Frank E. Kramer, Joan A. Sereno, and Lynne W. Shields have made contributions to this work. With regard to this application of the auditory-perceptual theory to the vowels, Hisao M. Chang and John W. Hawks deserve special mention, as together we collected data in our laboratory, organized data from the literature, and developed the I2 target zones as well as a variety of graphics routines for the appropriate visualization of the data. The detailed and constructive criticisms of the two reviewers, Bjorn Lindblom and an anonymous reviewer, and of Editor Joanne L. Miller, are greatly appreciated.

APPENDIX A: THE SENSORY REFERENCE

The concept of the sensory reference was developed with four goals in mind. (1) One goal was to correct the most glaring error of the formant-ratio theory of vowel quality, that is, the failure of this theory to distinguish certain vowels, such as /AA/ and /AO/ or /UH/ and /UW/, as illustrated by the work of Potter and Steinberg (1950). (2) Another goal was to maintain the effective elimination of the effects of talker age and sex achieved by formant-ratio theory. (3) A third goal was to introduce a mechanism whereby voice pitch could, under certain special circumstances, influence vowel identification, as indicated in the literature (Fant *et al.*, 1974; Fujisaki and Kawashima, 1968; Miller, 1953; and Scott, 1976). (4) The fourth goal was to introduce a mechanism whereby pitch modulations of certain frequencies could influence sensory and perceptual paths, and thus influence the segmentation process as well as vowel identity.

The approach to the first goal was to establish an absolute reference point against which the positions of the first three formants are to be judged. This was accomplished [in Eq. (6)] by defining the variable y in the auditory-perceptual space by

$$y = \log(SF1/SR). \tag{A1}$$

TABLE AI. Normalization of average ratios of $F1$ to $F0$.[a]

	Men	Women	Children
$GMF0$	132	223	264
$GMF1$	479	545	637
$\log[(GMF1)/(GMF0)^{1/3}]$	1.97	1.95	2.00
$y = \log(GMF1/SR)$	0.49	0.47	0.52

[a]Data are from Peterson and Barney (1952).

In this way, the absolute location of $SF1$ becomes a significant factor in the description of the spectrum as a locus in the auditory-perceptual space. Since, as a rule of thumb, the author used 125 Hz as an average pitch of the adult male voice and 225 Hz as an average pitch of the adult female voice, he selected their geometric mean, 168 Hz, as the reference frequency. Of course, better estimates of the mean pitch of all human speakers might lead to a better reference frequency, but its value is surely bounded between 150 and 200 Hz.

The second goal was to eliminate or reduce the effects of talker age and sex on the variable y without losing the advantage of a reference point. The solution was to allow the reference frequency of 168 Hz to be shifted, depending on the talker's vocal characteristics. As an initial step in this direction, the talker's mean pitch was chosen as a measure of his or her vocal characteristics. This variable was selected since it brought pitch in as a defining characteristic of the spectral pattern and gave pitch a role in defining a path through the auditory-perceptual space.

The next step was to define the sensory reference (SR) to be a function of the geometric mean of the talker's fundamental frequency, such that the average ratio ($SF1/SR$) across all glottal-source spectra would be independent of the talker's age and sex. To estimate the form of the desired function, the geometric mean of the first formants ($GMF1$) and fundamental frequencies ($GMF0$) reported by Peterson and Barney (1952) were separately calculated for children, women, and men. It was then found that the equation

$$k = \log[(GMF1)/(GMF0)^{1/3}] \tag{A2}$$

results in a value of k that is nearly constant across talkers, as shown in Table AI. It can be seen that k has a value of about 1.97, which is nearly independent of talker sex or age. It is not claimed that the proposed function is proven to be optimal; rather, it is claimed to be a simple, working approximation to a desired function that should eliminate all differences in k between talkers.

Now, an equation was needed that would meet the requirements that the sensory reference always be defined and that it quickly converges to a shifted value as information about the current talker's voice pitch becomes available. This is done [in Eq. (5)] by letting

$$SR = 168\ (GMF0/168)^{1.3}. \tag{A3}$$

By this formulation, the value of SR is shifted from the absolute value of 168 to a value appropriate to the current talker, such that average value y in Eq. (A1) is nearly identical across talker groups, as shown in the last row of Table AI. Since each talker's average pitch is fairly stable, the vow-

els /AA/ and /AO/ or /UH/ and /UW/ can be distinguished along the y dimension even when they have similar values of x and z. Furthermore, in special cases where changes in y can shift a vowel across the boundary of a vowel target zone, the identification of the vowel can be influenced by voice pitch. Thus the sensory reference concept as presented seems to achieve the first three goals mentioned in the introduction to this appendix.

Our approach to the fourth goal is indicated by Eq. (20) given in the text; that is,

$$SR = 168(GMF0/168)^{1/3} + \mathrm{FIL}(\mathrm{mod}\, F0), \qquad \text{(A4)}$$

where FIL(mod $F0$) indicates modulations of $F0$ filtered to eliminate rapid changes at pitch onset and offset, as well as the very slow changes or pitch declinations that do not seem to carry information relevant to the allophonic structure of an utterance. As stated in the text, this last term, FIL(mod $F0$), has yet to be evaluated in our research. Nonetheless, even though it seems likely that appropriate pitch modulations can serve to significantly alter sensory or perceptual paths, whether pitch modulations (and which ones) can influence the allophonic segments of speech are matters for future research.

Finally, it is noted that the sensory reference concept is similar to the more general adaptation level concept of Helson (1964). Helson's concept is that stimuli are judged in relation to an adaptation level or reference. The adaptation level is generally thought to be controlled by long-term factors, intermediate-term factors, and short-term factors. Long-term factors may include biological structure and global past experience; intermediate term factors may be factors common to the general environment in which judgments are being made; and short-term factors may be effects of preceding stimuli or sequences of stimuli. In the case of the sensory reference of the auditory-perceptual theory, the parameter of about 168 Hz is thought to be based on long-term factors; the shift produced by the current talker's average pitch is conceived of as an intermediate-term factor; and the possible effect of pitch modulation is thought of as a short-term factor. Viewed in this manner, the sensory reference of the auditory-perceptual theory is a special case of the adaptation level concept.

APPENDIX B: VOWEL DATABASE

The sources for the data used in the development of the I2 target zones are summarized in Table BI. The entry for each source indicates how many data points were converted into x, y, and z coordinates in the APS and whether the speakers were male (M), female (F), or children (C). The numbers before each source correspond to the number of the relevant comments that follow.

(1) These are the mean formant values for three groups of speakers (men, women, and children) from Table II (p. 183) in Peterson and Barney (1952).

(2) These are values taken from Tables I–III (pp. 17–19) in Peterson (1961).

(3) The values are from Table IV (p. 52) in Holbrook and Fairbanks (1962).

(4) These values are for one male speaker from Table VIII, list 1 (p. 21) in Lehiste (1962). The additional values for /ER/ are the averaged formant frequencies listed in Ta-

TABLE BI. Sources for the 435 data points used in the creation of the ten perceptual target zones.

Source	IY	IH	EH	AE	AA	AO	AH	UH	UW	ER	Total
(1) Peterson and Barney, 1952	1M 1F 1C	1M 1F 1C	1M 1F 1C	1M 1F 1C	1M 1F 1C	1M 1F 1C	1M 1F 1C	1M 1F 1C	1M 1F 1C	1M 1F 1C	30
(2) Peterson, 1961	4M 3F 3C	5M 2F 3C	5M 3F 3C	4M 2F 3C	4M 3F 3C	5M 3F 3C	5M 3F 3C	5M 2F 3C	3M 2F 3C	4M 2F 3C	99
(3) Holbrook and Fairbanks, 1962	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	10
(4) Lehiste, 1962	1M	1M	1M	1M	1M	1M	1M	1M	1M	8M	17
(5) Klatt, 1980	1	1	1	1	1	1	1	1	1	2	11
(6) CID corpus I	··· ···	16M 16F	16M 16F	··· ···	··· ···	··· ···	16M 16F	··· ···	··· ···	··· ···	96
(7) CID corpus II	··· ···	16M 16F	16M 16F	··· ···	··· ···	··· ···	16M 14F	··· ···	··· ···	··· ···	94
(8) CID corpus III	3M 3F	3M 3F	3M 3F	3M 3F	3M 3F	3M 3F	3M 3F	3M 3F	3M 3F	3F 3F	60
(9) CID corpus IV	1M 1F	1M 1F	1M 1F	1M 1F	1M 1F	1M 1F	1M 1F	1M 1F	1M 1F	··· ···	18
Totals	24	88	89	23	24	25	87	24	22	29	435

ble V (p. 65) for seven male speakers. For all tokens, $F0$ was set to 133 Hz.

(5) These proposed formant values for synthesis are from Table II (p. 986) in Klatt (1980), with $F0$ set equal to 133 Hz. All values are for nondiphthongized vowels. In addition, a second /ER/ token is based on the values for /R/ (Table III, p. 987).

(6) The CID recordings were made in an anechoic chamber using a high-quality microphone (B&K 4179/2660). All subjects were speakers of Midwestern American English. For this corpus (CID corpus I), two male and two female subjects were recorded reading lists of CVC words where C was either /D/ or /B/. The lists were read at both a normal and a fast rate of speech. The recordings were digitized at 20 kHz and spectral measurements were made at approximately the middle of the vocalic portion of the utterance (38%–50% from vowel onset). Spectral analysis was performed using the speech microscope described in Vemula *et al.* (1979).

(7) CID corpus II contains the same stimuli as CID corpus I, using four new speakers, two male and two female. Spectral analysis was done using ILS. LPC analysis was performed using a 25.6-ms Hamming window moving in 3.2-ms steps, a high-frequency preemphasis factor of 98%, and 24 poles. The fundamental frequency was determined using the cepstrally based algorithm provided by ILS. Formant values were picked automatically using the procedure described in Chang (1987a).

(8) CID corpus III is part of the data reported by Fourakis and Miller (1987). Two male and two female speakers were recorded producing the vowels in isolation and in sonorant context. Four tokens (two male, two female) of each vowel in isolation and two tokens (one male and one female) from sonorant context were included in this corpus. Spectral analysis was done using ILS. LPC analysis was performed using a 24-ms Hamming window moving in 1-ms steps, a high-frequency preemphasis factor of 98%, and 24 poles. Formant values of the vowel were converted into x, y, and z values by means of Eqs. (5)–(7). The values of x, y, and z were calculated and plotted in APS for each millisecond of the waveform, thus generating a sequence of points, or path, through the three-dimensional space. A portion of the path, selected by the experimenter, was taken as corresponding to the vowel. Factors influencing this selection were varied and included: (1) local extremes along the path, (2) regions of low velocity of movement along the path corresponding to "steady states," and (3) regions of high path curvature. Finally, the experimenter listened to the segmented portion to verify his selection. The experimenter averaged formant values over the selected part of the vowel path, thus computing one point per vowel token for each speaker.

(9) For CID corpus IV, two speakers were recorded reading the vowels in /B/–vowel–/B/ context. Spectral analysis was done using ILS. LPC analysis was performed using a 12.5-ms Hamming window moving in 2.5-ms steps. Formant values were found as described for corpus III, except that the experimenter picked a single point along the path to characterize the vowel and did not attempt to listen to the segmented point.

[1]The ARPABET notation is used in this paper (Lea, 1980). Correspondences with IPA of symbols used in this paper are: /IY/—/i/; /IH/—/ɪ/; /EH/—/ɛ/; /AE/—/æ/; /AA/—/ɑ/; /AO/—/ɔ/; /AH/—/ʌ̃/; /UH/—/ʊ/; /UW/—/u/; /ER/—/ɝ/; /L/—/l/; /D/—/d/; /B/—/b/; and /R/—/r/.

[2]The concept of the auditory-perceptual space is that spectral shapes can be characterized by a few, perhaps three, dimensions. Similar spaces have been suggested previously by authors such as Peterson (1952), Shepard (1972), and Pols (1977). There exist numerous possibilities for deriving these dimensions. For purposes of initiating exploration of this concept, in our work we have based our dimensions on the locations of the center frequencies of formants as described in the text. Clearly, spectral shapes are highly correlated with the locations of formant frequencies and, therefore, such locations offer an attractive approach to the specification of spectral shape. However, it would not be surprising if related metrics based on the entire spectral shape will be required in the future as it is well known that "formant tracking" in continuous speech is quite difficult. For the present, we carry out the required formant tracking using a variety of software aids and hand intervention. At the same time, we are trying to develop formant tracking algorithms that deal with problems of weak, merged, and missing formants as well as a variety of other problems mentioned in the work of Bladon (1982), McCandless (1974), Miller (1984a), and others.

[3]The z' axis of slab coordinates is perpendicular to the vowel slab. Thus, by Eq. (8), the center of the vowel slab has a z' of 0.704 log units (0.5772×1.22) and the thickness of the vowel slab is 0.156 log units (0.5772×0.27).

Ades, A. E. (**1977**). "Theoretical notes: Vowels, consonants, speech, and nonspeech," Psychoanal. Rev. **84**, 524–530.

Bell, A. G. (**1879**). "Vowel theories," Am. J. Otol. **1**, 163–180.

Bladon, A. (**1982**). "Arguments against formants in the auditory representation of speech," in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström (Elsevier Biomedical, Amsterdam, The Netherlands), pp. 95–102.

Broad, D. J. (**1976**). "Toward defining acoustic phonetic equivalence for vowels," Phonetica **33**, 401–424.

Broad, D. J., and Wakita, H. (**1977**). "Piecewise-planar representation of vowel formant frequencies," J. Acoust. Soc. Am. **62**, 1467–1473.

Chang, H. M. (**1987a**). "SWIS: See what I say; A speaker-independent word recognition system by phoneme-oriented mapping on a phonetically-encoded auditory-perceptual speech map," unpublished Ph.D. dissertation, Washington University, St. Louis, MO.

Chang, H. M. (**1987b**). "Automatic detection for diphthongs," J. Acoust. Soc. Am. Suppl. 1 **82**, S37.

Chiba, T. and Kajiyama, M. (**1941**). *The Vowel. Its Nature and Structure* (Tokyo-Kaiseikan, Tokyo, Japan).

Cooper, W. E., and Sorensen, J. M. (**1981**). *Fundamental Frequency in Sentence Production* (Springer, New York).

Cowan, N., and Morse, P. A. (**1986**). "The use of auditory and phonetic memory in vowel discrimination," J. Acoust. Soc. Am. **79**, 500–507.

Daniloff, R. G., Shriner, T. H., and Zemlin, W. R. (**1968**). "Intelligibility of vowels altered in duration and frequency," J. Acoust. Soc. Am. **44**, 700–707.

Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton, 's-Gravenhage, The Netherlands).

Fant, G. (**1973**). *Speech Sounds and Features* (MIT, Cambridge, MA).

Fant, G., Carlson, R., and Granström, B. (**1974**). "The [e]–[ø] ambiguity," *Speech Communication Seminar*, Vol. 3, Speech Transmission Laboratory, pp. 117–121.

Fletcher, H. (**1929**). *Speech and Hearing* (Van Nostrand, New York).

Fourakis, M., and Miller, J. D. (**1987**). "Measurements of vowels in isolation and in sonorant context," J. Acoust. Soc. Am. Suppl. 1 **81**, S17.

Fox, R. A. (**1985**). "Within- and between-series contrast in vowel identification: Full-vowel versus single-formant anchors," Percept. Psychophys. **3**, 223–226.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M., (**1962**). "The identification and discrimination of synthetic vowels," Lang. Speech **5**, 171–189.

Fujimura, O. (**1962**). "Analysis of nasal consonants," J. Acoust. Soc. Am. **34**, 1865–1875.

Fujisaki, H., and Kawashima, T. (**1968**). "The roles of pitch and higher formants in the perception of vowels," IEEE Trans. Audio Electroa-

coust. AV-16, 73–77.
Goldberg, R. F. (1986). "Perceptual anchors in vowel and consonant continua," M.S.E.E. thesis, MIT, Cambridge, MA.
Gottfried, T. L., and Strange, W. (1980). "Identification of coarticulated vowels," J. Acoust. Soc. Am. 68, 1626–1635.
Gottfried, T. L., Jenkins, J. J., and Strange, W. (1985). "Categorical discrimination of vowels produced in syllable context and in isolation," Bull. Psychon. Soc. 23, 101–104.
Greenwood, D. D. (1961). "Auditory masking and the critical band," J. Acoust. Soc. Am. 33, 484–501.
Harris, J. D. (1960). "Scaling of pitch intervals," J. Acoust. Soc. Am. 32, 1575–1581.
Helson, H. (1964). *Adaptation-level Theory* (Harper and Row, New York).
International Phonetic Association (1949). *The Principles of the International Phonetic Association* (International Phonetic Association, University College, London, England).
Iri, M. (1959). "Mathematical methods in phonetics," Gengo Kenkyu (Language Research) 35, 23–30.
Jones, D. (1914). *The Pronunciation of English* (Cambridge U. P., Cambridge, England).
Jones, D. (1919). "Experimental phonetics and its utility to the linguist," Proc. R. Inst. G. B. 22, 8–21.
Jongman, A., and Fourakis, M. (1987). "A cross-language study of vowel spaces and interference," J. Acoust. Soc. Am. Suppl. 1 81, S66.
Kent, R. D. (1979). "Isovowel lines for the evaluation of vowel formant structure in speech disorders," J. Speech Hear. Disord. 44, 513–521.
Klumpp, R. B., and Webster, J. C. (1961). "Intelligibility of time-compressed speech," J. Acoust. Soc. Am. 33, 265–267.
Koenig, W. (1949). "A new frequency scale for acoustic measurements," Bell Labs Rec. 27, 299–301.
Kurtzrock, G. (1956). "The effects of time and frequency distortion upon word intelligibility," unpublished Ph.D. dissertation, University of Illinois, Urbana, IL.
Lea, W. A. (1980). *Trends in Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ), p. 127.
Lloyd, R. J. (1890a). *Some Researches into the Nature of Vowel-Sound* (Turner and Dunnett, Liverpool, England).
Lloyd, R. J. (1890b). "Speech sounds: Their nature and causation (I)," Phonetische Studien 3, 251–278.
Lloyd, R. J. (1891). "Speech sounds: Their nature and causation (II–IV)," Phonetische Studien 4, 37–67, 183–214, 275–306.
Lloyd, R. J. (1892). "Speech sounds: Their nature and causation (V–VII)," Phonetische Studien 5, 1–32, 129–141, 263–271.
Lloyd, R. J. (1894). "Reply to criticisms by Dr. Pipping." Z. französische Sprache und Literatur 16, 201–206.
Macchi, M. J. (1980). "Identification of vowels spoken in isolation versus vowels spoken in consonantal context," J. Acoust. Soc. Am. 68, 1636–1642.
Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (1977). "The psychophysics of categorical perception," Psychol. Rev. 84, 452–471.
Macmillan, N. A., Braida, L. D., and Goldberg, R. F. (1987). "Central and peripheral processes in the perception of speech and nonspeech sounds," in *The Psychophysics of Speech Perception*, edited by M.E.H. Schouten (Martinus Nijhoff, Dordrecht, The Netherlands), pp. 28–45.
McCandless, S. S. (1974). "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. Acoust. Speech Signal Process. 22, 135–141.
Millar, J. B., and Ainsworth, W. A. (1972). "Identification of synthetic isolated vowels and vowels in H–D context," Acustica 27, 278–282.
Miller, J. D. (1984a). "Auditory processing of the acoustic patterns of speech," Arch. Otolaryngol. 110, 154–159.
Miller, J. D. (1984b). "Implications of the auditory-perceptual theory of phonetic recognition by the hearing impaired," ASHA Rep. #14, 45–48.
Miller, J. D. (1987a). "Auditory-perceptual processing of speech waveforms," in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson (Erlbaum, Hillsdale, NJ), pp. 257–266.
Miller, J. D. (1987b). "The auditory-perceptual theory of phonetic recognition: A synopsis" (unpublished).
Miller, J. D. (1987c). "Classification of vowel productions by means of perceptual target zones: A response to Ladefoged and Studdert-Kennedy," J. Acoust. Soc. Am. Suppl. 1 82, S82.
Miller, J. D., Engebretson, A. M., and Vemula, N. R. (1980). "Vowel normalization: Differences between vowels spoken by children, women, and men," J. Acoust. Soc. Am. Suppl. 1 68, S33.
Miller, J. D., Engebretson, A. M., and Vemula, N. R. (1983). "Observations on the acoustic descriptions of vowels as spoken by children, women, and men" (unpublished).
Miller, J. D., Sadoff, S. J., and Veksler, M. R. (1988). "Sensory-perceptual transformations in speech analysis," J. Acoust. Soc. Am. Suppl. 1 83, S70.
Miller, R. L. (1953). "Auditory tests with synthetic vowels," J. Acoust. Soc. Am. 18, 114–121.
Minifie, F. D. (1973). "Speech acoustics," in *Normal Aspects of Speech, Hearing, and Language*, edited by F. D. Minifie, T. J. Hixon, and F. Williams (Prentice-Hall, Englewood Cliffs, NJ), pp. 235–284.
Morse, P. A., Kass, J. E., and Turkienicz, R. (1976). "Selective adaptation of vowels," Percept. Psychophys. 19(2), 137–143.
Nearey, T. M. (1989). "Static, dynamic, and relational factors in vowel perception," J. Acoust. Soc. Am. 85, 2088–2113.
Nearey, T. M., and Assmann, P. F. (1986). "Modeling the role of inherent spectral change in vowel identification," J. Acoust. Soc. Am. 80, 1297–308.
Ochai, Y., Saito, S., and Sakai, Y. (1955). "Articulation study of speech qualities in rotational synchronous distortion," Mem. Fac. Eng. Nagoya Univ. 7, 40–48.
Okamura, M. (1966). "Acoustical studies of Japanese vowels in children. The formant construction and the developmental process," Jpn. J. Otol. (Tokyo) 69, 1198–1214.
Peterson, G. E. (1952). "The information bearing elements of speech," J. Acoust. Soc. Am. 24, 629–637.
Peterson, G. E. (1961). "Parameters of vowel quality," J. Speech Hear. Res. 4, 10–29.
Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175–184.
Pipping, H. (1893). "Review and criticism of Lloyd—1890a,b, 1891, 1892." Z. französische Sprache und Literatur 15, 157–171.
Pisoni, D. B. (1973). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Percep. Psychophys. 13, 253–260.
Pols, L. C. W. (1977). "Spectral analysis and identification of Dutch vowels in monosyllabic words," Institute for Perception TNO, Soesterberg, The Netherlands.
Potter, R. K., and Steinberg, J. C. (1950). "Toward the specification of speech," J. Acoust. Soc. Am. 22, 807–820.
Repp, B. H., Healy, A. F., and Crowder, R. G. (1979). "Categories and context in the perception of isolated steady-state vowels," J. Exp. Psychol.: Hum. Percept. Perform. 5, 129–145.
Sawusch, J. R., and Nusbaum, H. C. (1979). "Contextual effects in vowel perception I: Anchor-induced effects," Percept. Psychophys. 4, 292–302.
Sawusch, J. R., Nusbaum, H. C., and Schwab, E. C. (1980). "Contextual effects in vowel perception II: Evidence for two processing mechanisms," Percept. Psychophys. 5, 421–434.
Scott, B. L. (1976). "Temporal factors in vowel perception," J. Acoust. Soc. Am. 60, 1354–1365.
Sekimoto, S. (1982). "Perceptual normalization of frequency scale," Annu. Bull. RILP 16, 95–101.
Sharf, B. (1978). "Loudness," in *Handbook of Perception: Hearing*, edited by E. Carterette and M. Friedman (Academic, New York), pp. 187–242.
Shepard, R. N. (1972). "Psychological representation of speech sounds," in *Human Communication: A Unified View*, edited by P. B. Denes and E. E. David, Jr. (McGraw-Hill, New York).
Slawson, A. W. (1968). "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency," J. Acoust. Soc. Am. 43, 87–101.
Stevens, K. N., Fant, G., and Hawkins, S. (1987). "Some acoustical and perceptual correlates of nasal vowels," in *In Honor of Ilse Lehiste: Ilse Lehiste Pühendusteos*, edited by R. Channon and L. Shockey (Foris, The Netherlands), pp. 241–254.
Strange, W. (1989). "Evolving theories of vowel perception," J. Acoust. Soc. Am. 85, 2081–2087.
Syrdal, A. K., and Gopal, H. S. (1986). "A perceptual model of vowel recognition based on auditory representation of American-English vowels," J. Acoust. Soc. Am. 79, 1086–1100.
Tiffany, W. R., and Bennett, D. N. (1961). "Intelligibility of slow-played speech," J. Speech Hear. Res. 4, 248–258.
Vemula, N. R., Engebretson, A. M., Monsen, R. B., and Lauter, J. L. (1979). "A speech microscope," in *Speech Communication Papers*, edited by J. J. Wolf and D. H. Klatt (Acoustical Society of America, New York), pp. 71–74.

Wier, C. C., Jesteadt, W., and Green, D. M. (1977). "Frequency discrimination as a function of frequency and sensation level," J. Acoust. Soc. Am. **61**, 178–184.

Zwicker, E., and Terhardt, E. (1980). "Analytical expressions for critical band rate and critical bandwidths as a function of frequency," J. Acoust. Soc. Am. **68**, 1523–1525.

Zwislocki, J. J. (1978). "Masking: Experimental and theoretical aspects of simultaneous, forward, backward, and central masking," in *Handbook of Perception: Hearing*, edited by E. Carterette and M. Friedman (Academic, New York), pp. 283–336.

LANGUAGE AND SPEECH, 1989, 32(3), 221 – 248

THE ACOUSTIC VOWEL SPACE OF MODERN GREEK AND GERMAN*

AFOSR Grant 6-AFOSR-86-0335
Final Technical Report
Appendix

ALLARD JONGMAN,
MARIOS FOURAKIS
and
JOAN A. SERENO
Central Institute for the Deaf, St. Louis

The spectral characteristics of vowels in Modern Greek and German were examined. Four speakers of Modern Greek and three speakers of German produced four repetitions of words containing each vowel of their native language. Measurements of the fundamental frequency and the first three formants were made for each vowel token. These measurements were then transformed into log frequency ratios and plotted as points in the three-dimensional auditory-perceptual space proposed by Miller (1989). Each vowel token was thus represented by one point, and the points corresponding to each vowel category were enclosed in three-dimensional target zones. For the present corpus, these zones differentiate the five vowels of Modern Greek with 100% accuracy, and the fourteen vowels of German with 94% accuracy. Implications for the distribution of common vowels across languages as a function of vowel density are discussed.

Key words: vowel space, Greek, German

INTRODUCTION

The process of defining a vowel space for a language, or a universal space for all languages, can be characterized as having three facets corresponding to three different stages in the communication process (Lindblom, 1986). The articulatory stage, at which the vocal tract is shaped so as to produce the intended vowel, defines an articulatory vowel space by the possible positions of the tongue and the jaw, and by the shape of the lips. The acoustic stage, at which the sound radiating from the lips propagates through air, defines an acoustic space by the relative distribution of energy in the time and frequency domains. Finally, the auditory stage, at which the sound is processed by the

---

* The first and third authors are presently at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. A partial report of the results was presented at the 113th Meeting of the Acoustical Society of America in Indianapolis, IN, May 11–15, 1987. The research was supported by NIH Grant ND21994 and AFOSR Grant 860335 to Central Institute for the Deaf. The authors wish to thank James D. Miller, John W. Hawks, Steven J. Sadoff, Frank E. Kramer, and Melissa P. Piasecki for invaluable assistance.

Address all correspondence to Allard Jongman, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands.

ear and perceived as a linguistic unit, defines an auditory vowel space.

A traditional problem has been to uniquely describe and distinguish phonemically different vowels within this articulatory, acoustically, and auditorily defined vowel space. Acoustically, it is often the case that vowel sounds perceived as representing the same phonemic category differ in terms of their spectral and/or temporal properties. Several sources of variability may account for these differences. The acoustic characteristics of a specific vowel can be affected by: (1) the sex, age, and size of a speaker (inter-speaker variation); (2) the phonetic context in which the vowel is produced (coarticulatory effects); (3) the rate of speech at the moment the vowel is produced; (4) the stress value assigned to the syllable that contains the vowel; and (5) intonation contour and phonation types. The last four sources collectively are usually referred to as intra-speaker variation. It is certainly a task for any theory of speech perception to account for the way in which a listener processes these differences during normal speech understanding. Both inter- and intra-speaker variations may cause instances of phonemically identical vowels to occupy different regions in the vowel space, or instances of phonemically different vowels to occupy overlapping or identical regions in the vowel space (Peterson and Barney, 1952).

Several normalization schemes have been proposed in the literature in order to map the vowels of a language onto unique, non-overlapping regions in the vowel space. Such normalization algorithms attempt to extract invariant features of vowel sounds by taking into account the frequency values of two or more formants and by transforming them in various ways including mel scales (Fant, 1973), deviations from log means of vowel formant frequencies (Nearey, 1978), Bark versus sones/Bark representations (Lindblom, 1986), Bark differences (Syrdal and Gopal, 1986), and multidimensional scaling techniques (e.g., Wright, 1986). The majority of such normalization schemes has focused on American English vowels. However, even within a single language, normalizing for variability due to speaker, phonetic context, and speaking rate proves to be quite difficult. Most recently, Miller (1987a; 1989), in an attempt to account for intra- and inter-speaker differences, has proposed a general framework for the representation of speech sounds in an "auditory-perceptual space" in terms of log ratios of F0, F1, F2, and F3. The present paper is an attempt to extend this framework, which was based on the speech sounds of American English, to other languages. In particular, we address the problem of mapping vowels onto unique regions in the vowel space for Modern Greek, a language with five vowels, and for German, a language with 15 vowels, within this general approach.

The vowel space in the auditory-perceptual theory

In the auditory-perceptual theory (Miller, 1989), speech sounds, vowels in this particular case, are mapped onto an auditory-perceptual space (APS) of three dimensions. The vowel space proposed by Miller can be viewed as an outgrowth of proposals found in earlier work by Peterson (1952), Shepard (1972), and Pols (1977). The dimensions of the auditory-perceptual space are defined by the short-term spectrum of a vowel sound in terms of log frequency ratios, and specifically by the positions of the first three prominences in the spectrum relative to each other and to a low frequency reference.

The following equations define the auditory-perceptual space:

$$x = \log(SF3/SF2) \qquad \text{Eq. 1}$$

$$y = \log(SF1/SR) \qquad \text{Eq. 2}$$

$$z = \log(SF2/SF1) \qquad \text{Eq. 3}$$

SF1, SF2, and SF3 represent the frequency locations of the first three significant prominences of the short term spectral envelope of the vowel waveform. SR is a reference frequency, which is shifted slightly by the average spectrum of the current speaker. This reference is calculated as follows:

$$SR = 168(GMF0/168)^{1/3} \qquad \text{Eq. 4}$$

where GMF0 is the geometric mean of the speaker's F0 for the utterance. The choice of these variables, and the motivation for the sensory reference, are discussed extensively in Miller (1989).

Miller (1989) presented formant values of American English vowels taken from the literature and from measurements in the C.I.D. laboratories. These values were converted into log frequency ratios, using Equations 1 through 4, and were then plotted in the three-dimensional space. The 406 data points representing the nine non-diphthongized vowels of English (/i, ɪ, ɛ, æ, a, ʌ, ɔ, ʊ, u/) are shown plotted in Figure 1 (Panel A: front view; Panel B: side view). As is clear from Panel B, these points fall in a narrow slab ('the vowel slab') of the three-dimensional auditory-perceptual space, with the width of the slab being quite small.

The points in the auditory-perceptual space corresponding to the same phonetic vowel category can be grouped accordingly into target zones. Figure 2 shows the nine non-overlapping vowel target zones of American English which can account for 93% of the data. (See Miller, 1989, for a complete listing of the individual data points and corresponding target zones.)

These target zones exhibit no overlap even though the data points that define them represent vowels spoken by speakers of different ages and gender, in various phonetic contexts, and at different speaking rates. For expository purposes, the axes of the APS in Figure 2 have been rotated by a transformation to bring the vowel slab to a vertical position using the following set of equations:

$$x' = 0.70711(y-x) \qquad \text{Eq. 5}$$

$$y' = 0.81622(z) - 0.4081(x+y) \qquad \text{Eq. 6}$$

$$z' = 0.5772(x+y+z) \qquad \text{Eq. 7}$$

This transformation does not translate the origin, but only rotates the axes. We refer to the coordinates x', y', and z' defined by Equations 5–7, as slab coordinates.

These data suggest that the representation of vowels by means of the logarithmic ratios of the first three formants and a reference frequency can uniquely characterize the vowels of American English. Such a representation may thus account for much of the variability due to phonetic context and variations in speaking rate, as well as for differences among speakers in terms of gender and age.

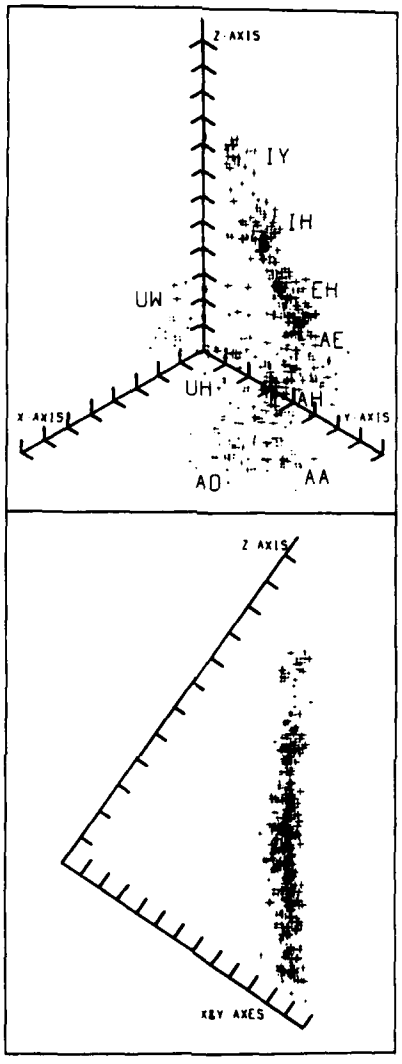


Fig. 1. Panel A. Formant measurements of 406 American English monophthongal vowel tokens transformed into points in the three-dimensional auditory-perceptual space. x, y, and z axes are in 0.1 log units, and the point of origin is (0, 0, 0). The symbols are in Arpabet notation: IY = [i], IH = [ɪ], EH = [ɛ], AE = [æ], AO = [ɔ], AA = [a], AH = [ʌ], UW = [u], UH = [ʊ].

Panel B. Side view of the APS showing how all vowel tokens fall into a narrow slab of the APS (the 'vowel slab'). In this orientation, the x and y axes appear to fall on top of each other. (Figure taken from Figure 13 in "Auditory-perceptual interpretation of the vowel", by J.D. Miller, *Journal of the Acoustical Society of America*, **85**, 2114–2134. Copyright 1989 by Acoustical Society of America. Used by permission.)

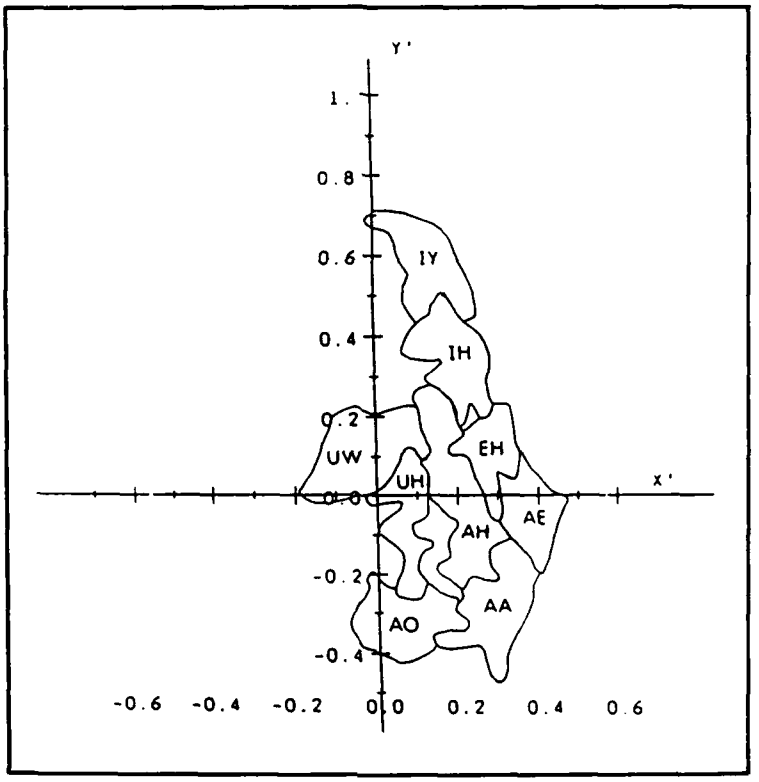


Fig. 2. Target zones for the nine monophthongal vowels of American English in slab coordinates (see Equations 5–7). In this front view, the z' axis is perpendicular to the x'y' plane. x', y', and z' axes are in 0.1 log units, and the point of origin is (0, 0, 0). (Figure based on Figure 14 in "Auditory-perceptual interpretation of the vowel", by J.D. Miller, *Journal of the Acoustical Society of America*, **85**, 2114–2134. Copyright 1989 by Acoustical Society of America. Used by permission.)

A cross-language evaluation of the auditory-perceptual theory

The auditory-perceptual theory of speech perception was initially developed for, and applied to, the vowel sounds of American English. However, for the theory to be universal, an important challenge is to see how it accommodates vowel inventories that differ from those of English. For example, it is of interest to see how the vowels of languages with smaller vowel inventories are organized in the auditory-perceptual space. More importantly, the auditory-perceptual theory must also be able to account for more complex inventories. In a first attempt, the present study examines the vowel spaces of Modern Greek and German. Modern Greek has a simple five-vowel system, consisting of [i, e, a, o, u] (Householder, Kazazis, and Koutsoudas, 1964). German, on the other hand, has a complex 15 vowel inventory with both tense and lax vowels, and rounded and

unrounded vowels, consisting of [i:, i, ɛ, e:, e, y:, y, ø:, ø, a:, a, o:, o, u:, u][1] (Moulton, 1962; Jørgensen, 1969).

To our knowledge, no studies of the spectral characteristics of Greek vowels have been reported. For German, two studies were of interest for the purpose of the present analysis. In a spectrographic study, Jørgensen (1969) reported the formant values of the 15 German vowels as produced by six male speakers. In addition, these vowels were plotted in a linear F1 by F2 space for each speaker individually. In general, Jørgensen found a tendency for the lax vowels to be more centrally located than their tense counterparts. There was, however, considerable overlap among vowels even for a single speaker, and superimposing the data plots for all six speakers showed extensive overlap for some vowels. Unfortunately, Jørgensen did not include a plot or any statistical evaluation for the combined data of all six speakers. In a recent study, Iivonen (1987) reported formant measurements of the 15 German vowels produced by one male speaker (with exceptionally high F0, around 200 Hz). This study was not designed to uniquely characterize the German vowels; it primarily compared the accuracy of three methods of spectral analysis.

Given the lack of any data on the Greek vowels, and the scarceness of data on the German vowels, the present study serves two purposes. First, it provides basic vowel measurements for Greek and German by listing values of fundamental frequency and the first three formants for each vowel. Second, this study tries to uniquely characterize these vowels by converting the formant values into log frequency ratios and plotting them as points in the three-dimensional auditory-perceptual space.

METHOD

Subjects

Subjects were recruited from the student body and faculty of Washington University, St. Louis, MO, with the exception of one Greek speaker, who was visiting the United States. Recruitment of native speakers who had only recently arrived in the United States drastically restricted the subject pool. The Greek subjects were two graduate students, one faculty member, and one visiting scholar, all native male speakers of Modern Greek (three from Athens, one from Patras). All four subjects had been in the United States between 10 days and four months at the time of the recordings. The German speakers were three female graduate students, two from Cologne and one from Bonn, all native speakers of German. All three speakers can be classified as speaking the 'Central Franconian' dialect (Wiesinger, 1983). At the time of the recordings, they had been in the United States for a duration of two to six months.

---

[1] We have adopted the phonetic symbols used by Jørgensen (1969) so that his results can be directly compared to the present results.

TABLE 1

Greek and German test words containing the 5 vowels of Greek and 14 vowels of German, respectively. IPA symbols representing the vowel of the first syllable of each word are given to the right of each column.

Modern Greek	Gloss	IPA	German	Gloss	IPA
Pita	pie	[i]	Stiele	handles	[i:]
Peta	fly!	[e]	Stille	silence	[i]
Pata	step!	[a]	stehle	(I) steal	[e:]
Pote	when	[o]	Stelle	place	[e]
Puse	where are you?	[u]	fühle	(I) feel	[y:]
			fülle	(I) fill	[y]
			Höhle	cave	[ϕ:]
			Hölle	hell	[ϕ]
			Buhle	lover	[u:]
			Bulle	bull	[u]
			Sohle	sole	[o:]
			solle	(I) ought to	[o]
			fahle	pale (pl.)	[a:]
			Falle	trap	[a]

Materials

Two word lists (see Table 1) were used: one for the Greek and one for the German speakers. For Greek, stimulus words were selected such that the five target vowels appeared in approximately the same context. For German, stimuli were taken from Moulton (1962) such that words containing a tense-lax vowel opposition formed minimal pairs (e.g., 'stehle' – 'Stelle'); the following consonant was always /l/. The low, front long vowel [ɛ] was excluded, since it is subject to much dialectal variation (Moulton, 1962; Jørgensen, 1969). Thus, the number of German vowels in the present study was 14. Randomized lists were made that contained four repetitions of each word. The randomized Greek and German words were placed in the following carrier sentences, respectively:

Greek:	[θa po ——ksana].	"I will say——again."
German:	Ich sage —— noch einmal.	"I say—— one more time."

The subjects were instructed to familiarize themselves with the list of sentences before the recordings were made. The total number of tokens recorded was 80 (5 vowels X 4 speakers X 4 repetitions) for Greek, and 168 (14 vowels X 3 speakers X 4 repetitions) for German.

Recording procedure

Subjects were recorded in an anechoic chamber using a special, low-noise microphone/preamplifier combination (Bruel and Kjaer 4179/2660). The microphone was placed at a height equal to, and 0.5 m in front of, the speaker's mouth (0° angle of incidence). Conversational speech levels were used. The microphone output was fed directly to a Sony PCM-501ES digital audio recorder (16 bit mode) with a JVC 720 VCR serving as the storage medium. A reading timer device, designed and built in-house, was used to regulate subjects' speed for recitation of the sentences.

Analysis and further processing

The recordings were digitized at 20 kHz with a 10-kHz low-pass filter setting with 16 bit precision and stored as files to be processed by the commercial software package ILS (Interactive Laboratory System). The primary sampled data files were then high-pass filtered at 50 Hz to remove any incidental low frequency noise. First, LPC analysis was performed using a 24 msec Hamming window moving in 1 msec steps, a preemphasis factor of 98%, and 24 poles. In addition, a cepstrally based algorithm was used to determine the fundamental frequency. Next, a frequency spectrum was derived at each millisecond of signal by means of a Fast Fourier Transform of the coefficients resulting from the linear prediction analysis. A program written in-house was used to extract the F0, F1, F2, and F3 values from the ILS secondary file and store them in a table-format file along with the corresponding frame (msec) number. In the case of merged or missing formants, a root solving command was used to enhance the resonance calculations of the FFT.

The values of F0, F1, F2, and F3 were converted into x, y, and z coordinates by means of Equations 1–3. These x, y, and z values were calculated and plotted at each millisecond of the waveform, thus generating a sequence of points, or path, through the three-dimensional space. The distance between any two points in the path corresponded to the amount of spectral change that had taken place within that time period.

For each vowel, we selected the 'steady-state' portion in the following manner. When we displayed a path in the APS on a graphics screen (Evans and Sutherland PS300), we could determine the portion of the path exhibiting very little movement in the space, i.e., corresponding to slow or no spectral change. Using cursors, we then were able to determine the beginning and end of these portions. We extracted the corresponding segment of the original waveform, smoothed it on both sides using a 12 msec half Kaiser window, and listened to it to verify that it sounded like the intended vowel. This procedure of excising the 'steady-state' portion was applied to all vowel utterances of both groups of speakers.

Once the tokens were verified, we took an average of the x, y, and z coordinates over that part of the vowel showing little change, yielding one point per vowel token for

each speaker. Thus, each vowel token was represented by a single point in the APS, which was based on an average of the x, y, and z values corresponding to the 'steady-state' portion of the vowel and which had been perceptually verified to be the intended vowel.

Using the procedures outlined above, we collected 80 points for the Greek vowels, and 167 points for the German vowels. (One token of the German vowel [o:] was disqualified due to equipment malfunction during the recording session.) The Appendix lists all the absolute frequency values (F0, F1, F2, F3) for each token of each vowel for both the Greek and the German speakers. The values reported in the Appendix are geometric means of F0, F1, F2, and F3 over the 'steady-state' portion of the vowel tokens as defined above.

Data points from the literature

In order to expand our data base, we incorporated measurements taken from the available literature. For German, we converted Jørgensen's (1969) measurements into points in the APS by means of Equations 1–3, using a constant F0 of 135 Hz (approximating the average male pitch of the Peterson and Barney (1952) data base) for all six male speakers. Jørgensen did not report F3 values for 14 out of 24 instances of mid and high back vowels. For these cases, we used as an F3 value the geometric mean of all his reported F3 values for that vowel for all speakers. These procedures yielded 84 points for the German vowels.

In addition, we converted formant values reported by Iivonen (1987) into points in the APS. Iivonen included measurements for fundamental frequency and the three formant frequencies at three different locations in the vowel: One taken at the "target for F1", one at the "target for F2", and one at "maximum volume velocity" (Iivonen, 1987, p. 128). In converting these values to points in the APS, we used the geometric mean of these three measurements for each vowel. This yielded 14 more points for the German vowels.

In sum, the Greek data base consisted of 80 tokens based on our own measurements, while the German data base consisted of 265 vowel points, of which 167 were based on our own measurements, 84 were taken from Jørgensen (1969), and 14 were taken from Iivonen (1987).

RESULTS

Modern Greek

The vowel productions of the Greek speakers are shown plotted in the APS (front view) in Figure 3. Each symbol represents one vowel token by a speaker. The middle of each symbol represents the actual location of each token. Points belonging to the same phonetic category are enclosed in target zones. These target zones are, in fact, three-dimensional objects, the plane projections of which are shown in Figure 3.[2] All

[2] The target zones were drawn on a high-resolution color graphics terminal (Evans and Sutherland PS300) which emulates a third dimension. No constraints were placed on the shape of the target zones, except for the fact that they had to be continuous. All zones were drawn by hand, using a computer-aided method with a resolution of 0.01 log units. This resolution was chosen so that the present classification performance could be compared to that for American English as reported by Miller (1989), who used the same resolution to draw the target zones.

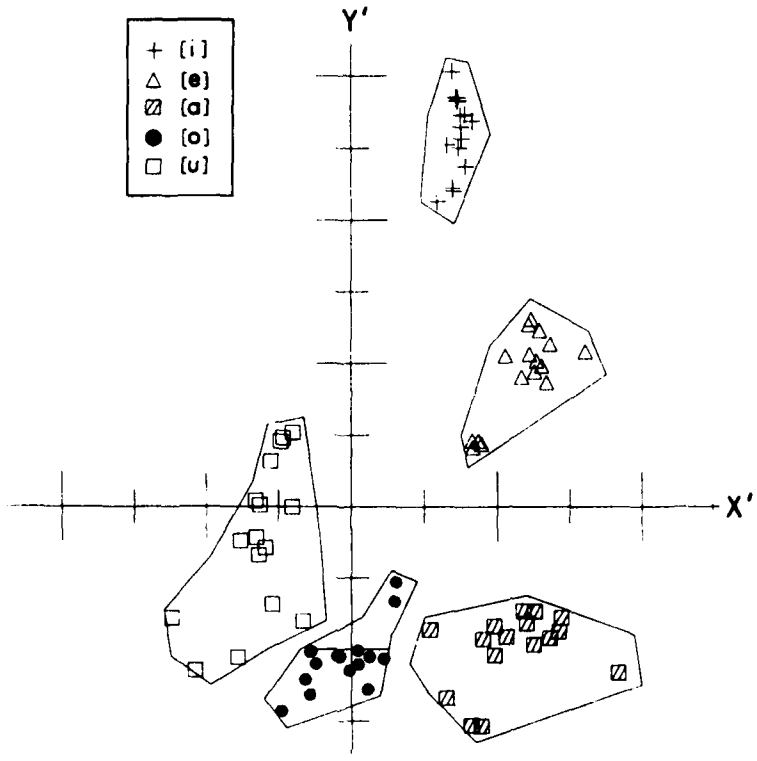


Fig. 3. Data points and target zones for the five Greek vowels shown in front view using slab coordinates. See Figure 2 for axis labels and units. Each symbol represents one vowel token produced by one speaker.

the present data points can be enclosed by distinct, non-overlapping zones. Thus, these zones enable us to describe the present corpus of Greek vowels with 100% accuracy.

It is remarkable that the target zones for [i] and [e], in particular, leave a lot of space unoccupied. One might expect much more variation in the vowel productions of Greek speakers since there are no competing phonemes in the area, as opposed to American English (see, for example, Lindblom, 1986). Of course, further analysis of vowels produced by both male and female speakers (since only male speakers were used in the present study) and in more phonetic contexts might result in target zones for Modern Greek which are less compact.

German

The German data present a more complex pattern. Our German corpus offers not only a greater vowel inventory, but also vowel tokens from both male and female speakers. The 265 points representing the German vowels were plotted in the APS. In an attempt to minimize the amount of overlap among the 14 vowels, each set of points

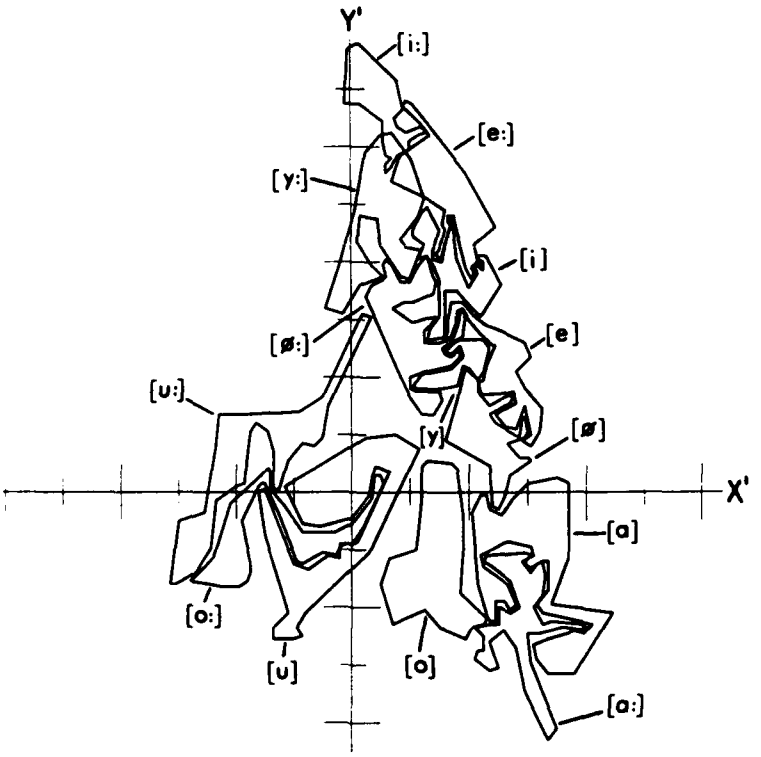


Fig. 4. Target zones for the 14 German vowels shown in front view using slab coordinates. See Figure 2 for axis labels and units. The zones associated with each of the vowels are indicated by the labels.

belonging to a particular vowel category was enclosed by an outline of a corresponding target zone.

Figure 4 shows a front view of the auditory-perceptual space with the target zones for the 14 German vowels.[3] As is apparent from Figure 4, these target zones are quite irregular in shape, thus accommodating some differences in speaker and phonetic context. For the present study, we have simply drawn the target zones such that they enclose as many appropriate data points as possible. However, with this relatively small sample, it remains an open question whether such zones will prove to have any explanatory value. In order to address this question, perceptual verification experiments must be conducted to establish the 'psychological reality' of the target zone boundaries. Such experiments are currently being conducted in our laboratories to evaluate the target zones established for American English. However, until this issue has been settled, we believe that there is no *a priori* reason to prefer smooth zones (e.g., circles or ellipses) over the present

[3] For the sake of clarity, individual data points are not shown in this Figure but are shown for each vowel in Figures 5, 6, and 7.

irregular zones.[4]

For the 14 German vowels, the three-dimensional irregular zones were able to describe the present corpus of German vowels with 94% accuracy. The zones group the data points according to phonetic category without any overlap. Since Figure 4 is a two-dimensional representation of the three-dimensional space, some target zones may appear to overlap. However, none of the target zones show any overlap when taking all three dimensions into account.

Table 2 summarizes the results of mapping the intended productions onto distinct target zones. The overall accuracy is approximately 94% correct. Only 16 vowel tokens out of a total of 265 tokens were misclassified, with the majority of those points falling in regions not belonging to any target zone. The vowel with the lowest score is [y] (53%), which exhibits considerably more scatter than any of the other vowels. With the exception of this vowel, the percent correct classifications range from 89% to 100%.

The individual target zones with their corresponding data points are shown in Figures 5, 6, and 7. Figure 5 shows a front view of the data points and target zones for the front unrounded vowels [i:, i, e:, e]. Figure 6 shows a front view of the data points and target zones for the front rounded vowels [y:, y, ø:, ø]. Figure 7 shows a front view of the data points and target zones for the back vowels [u:, u, o:, o, a:, a].

With respect to the front-back distinction, there is a clear separation between front vowels, on the one hand, and back vowels on the other. Front vowels occupy the APS quadrant with positive x' and y' values, while back vowels occupy quadrants with negative x' and either positive or negative y' values.

As for the unrounded-rounded distinction, the target zones for the front rounded vowels ([y:, y, ø:, ø]) fall behind those for the front unrounded ones ([i:, i, e:, e]),

---

[4] In addressing the issue of the irregular boundaries of the target zones, Miller (1987b) presented the 'third iteration' target zones for English, which classified 2051 vowel data points into the nine monophthongal categories of English with 95.8% accuracy. Of these 2051 points, 1420 were from Peterson and Barney (1952) and 631 were collected at Central Institute for the Deaf. In order to examine the effect of highly irregular borders on the percentage of correct classification, Miller (1987b) applied a smoothing algorithm to these target zones, which iteratively averaged the border lines every three points. He applied this smoothing 50, 100, 400, and 1000 times. After 100 times, for example, the target zone for [ʌ], which is highly irregular since in APS it is flanked by many adjacent target zones, became a three-dimensional oval-shaped object with no appendages, about 0.25 log units long and 0.18 log units wide in front view. The correct classifications for each of the smoothings are given in the following table as a percentage of the total 2051 points:

Smoothing iterations	0	50	100	400
Correct classification (%)	95.8	72.5	69.5	45.1

After about 1000 iterations, the zones collapsed into single points and the classification percentage approached zero. We did not apply this algorithm to the German target zones because of the small number of data points. However, we are currently in the process of extending our German database and we hope to be able to refine the German target zones as well as get some estimate of the effect of irregular boundaries using perception experiments.

TABLE 2

Classification of German vowels by means of perceptual target zones. The rows show the intended vowels and the columns the target zones onto which they map

Produced vowel	Mapped onto: i:	i	e:	e	y:	y	ɸ:	ɸ	u:	u	o:	o	a:	a	Unclaimed areas	% correct
i:	19	–	–	–	–	–	–	–	–	–	–	–	–	–	–	100
i	–	17	–	–	1	–	–	–	–	–	–	–	–	–	1	89
e:	–	–	19	–	–	–	–	–	–	–	–	–	–	–	–	100
e	–	–	–	19	–	–	–	–	–	–	–	–	–	–		100
y:	–	–	–	–	19	–	–	–	–	–	–	–	–	–		100
y	–	2	–	–	–	10	1	–	–	–	–	–	–	–	6	53
ɸ:	–	–	–	–	–	1	18	–	–	–	–	–	–	–	–	95
ɸ	–	–	–	–	–	–	–	19	–	–	–	–	–	–	–	100
u:	–	–	–	–	–	–	–	–	18	–	–	–	–	–	1	95
u	–	–	–	–	–	–	–	–	–	18	–	–	–	–	1	95
o:	–	–	–	–	–	–	–	–	–	–	17	–	–	–	1	94
o	–	–	–	–	–	–	–	–	–	–	–	19	–	–	–	100
a:	–	–	–	–	–	–	–	–	–	–	–	–	19	–	–	100
a	–	–	–	–	–	–	–	–	–	–	–	–	–	18	1	95

Total correct: 249/265 = 94

as shown in Figure 8. A comparison of the front unrounded and front rounded vowels in terms of z' values revealed that the front unrounded vowels have a mean z' value of approximately 0.70 log units, whereas the front rounded vowels are further back with a mean z' value of approximately 0.65 log units. The distinction between the front rounded and front unrounded vowels in terms of z' values eliminates the apparent overlap in the two-dimensional front view among the vowels [i:], [e:], and [y:]. As shown in Figure 4, the target zone for rounded [y:] seems to overlap with those for unrounded [i:] and [e:]. However, when the vowel slab is rotated 90 degrees into side view, it can be seen that these target zones do not overlap. Figure 9 shows such a side view of the data

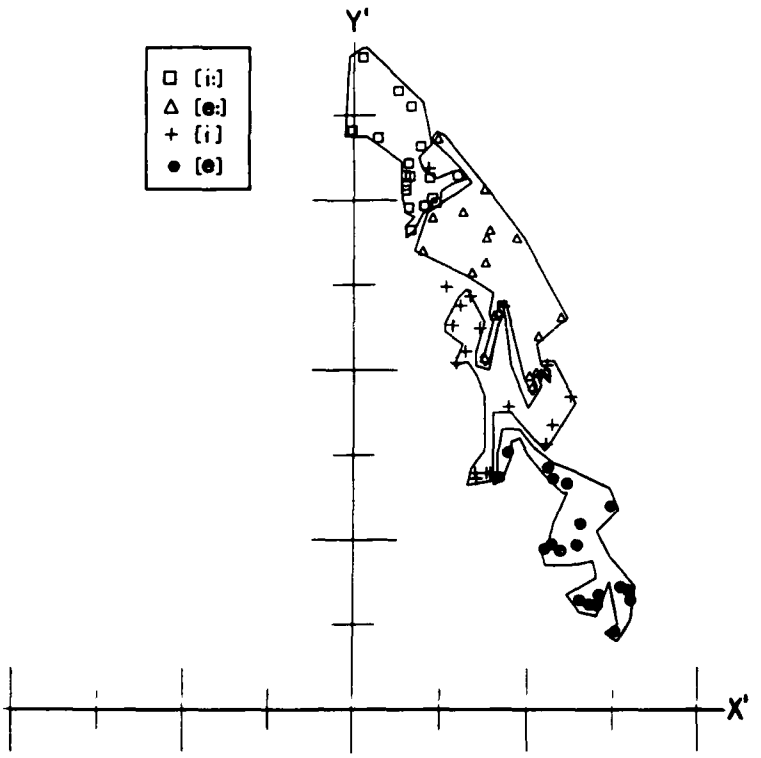


Fig. 5. Data points and target zones for the German front unrounded vowels ([i:, i, e:, e]) shown in front view using slab coordinates. See Figure 2 for axis labels and units. Each symbol represents one vowel token produced by one speaker.

points for [i:] and [e:] (which overlap in side view only) and the points representing [y:]. As can be seen, all [y:] points are behind the data points for [i:] and [e:]. Thus, those unrounded-rounded vowel pairs that overlap in the front view can be seen to be nonoverlapping in the side view.

Finally, German tense and lax vowels map onto distinct target zones, with target zones for tense vowels generally having larger y' values than those for their lax counterparts. Since the value of y' is largely dependent on the value of z (see Equation 6), this finding is in accord with the traditional observation (e.g., Jørgensen, 1969) that tense (long) vowels tend to have a lower F1 (i.e., are less centralized) than their lax (short) counterparts.

In order to compare the performance of the APS to that of other classification schemes, we conducted linear discriminant analyses using combinations of F0, F1, F2, and F3, as well as the APS parameters x, y, and z, as the classificatory variables.[5] Linear

[5] We thank one of the reviewers, Terry Nearey, for making his linear discriminant analysis program available to us.

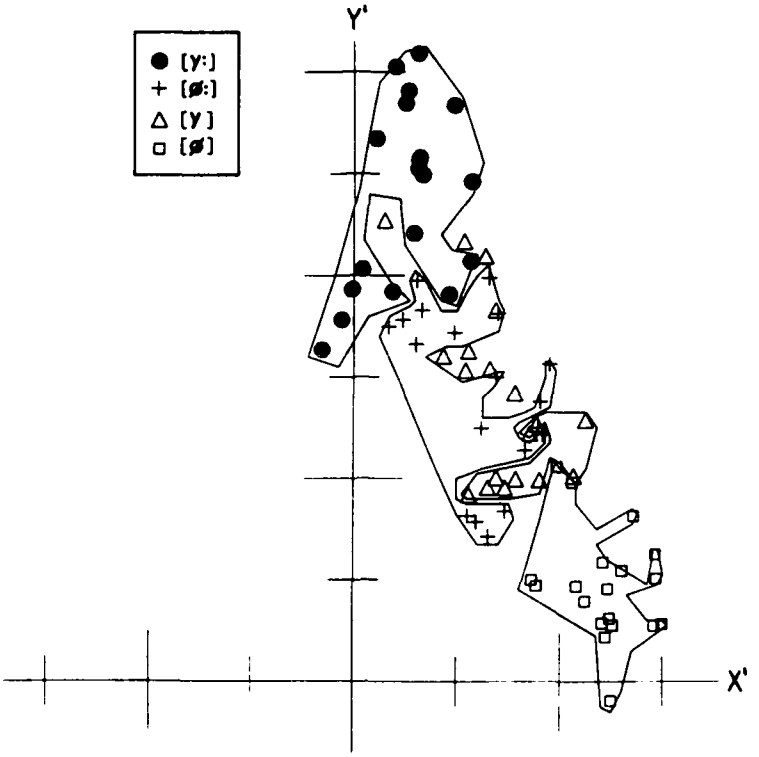


Fig. 6. Data points and target zones for the German front rounded vowels ([y:, y, ø:, ø]) shown in front view using slab coordinates. See Figure 2 for axis labels and units. Each symbol represents one vowel token produced by one speaker.

discriminant analysis is a technique to verify that apparent clusters are real, and to decide to which cluster a new data point should be assigned. The discriminant analysis method as applied to similar issues is described in Assmann, Nearey, and Hogan (1982), Syrdal (1985), and Syrdal and Gopal (1986). Briefly, this analysis uses group means to assign each individual token to a group on the basis of the estimated *a posteriori* probability of group membership. The results presented here are from the R (resubstitution) method of classification, which provides an index of the resolution into groups of the present data set. In addition, the *a posteriori* probability (APP) indicates the relative strength of group membership. For moderate sample sizes such as the present one, this index may be more informative than percent correct classification (Assmann *et al.*, 1982).

The results of the discriminant analyses for the variables F0, F1, F2, and F3, as well as x, y, and z, are shown in Table 3. It can be seen that various combinations of these variables produce reasonably similar results. However, the addition of F0 to either the first two or the first three formants improves resolution, increasing the APP to 0.58 and 0.60, respectively. The x, y, z coordinate system yields a very similar performance,

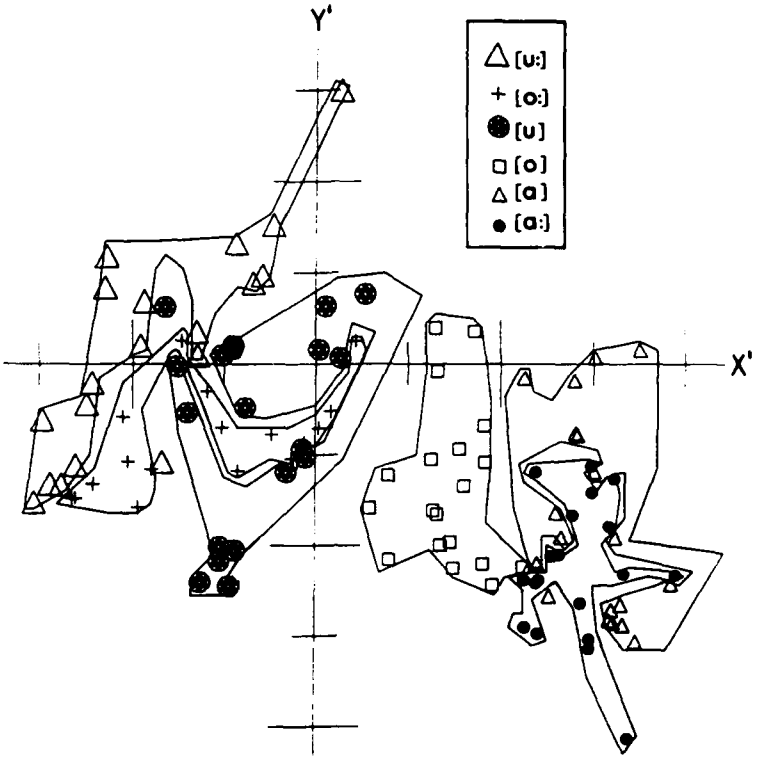


Fig. 7. Data points and target zones for the German back vowels ([u:, u, o:, o, a:, a]) shown in front view using slab coordinates. See Figure 2 for axis labels and units. Each symbol represents one vowel token produced by one speaker.

TABLE 3

Results of discriminant analysis for German vowels

Variables	% correct	*a posteriori* probability (APP)
F1, F2	63.4	0.50
F1, F2, F3	61.9	0.51
F0, F1, F2	67.6	0.58
F0, F1, F2, F3	67.2	0.60
x, y, z	65.3	0.58

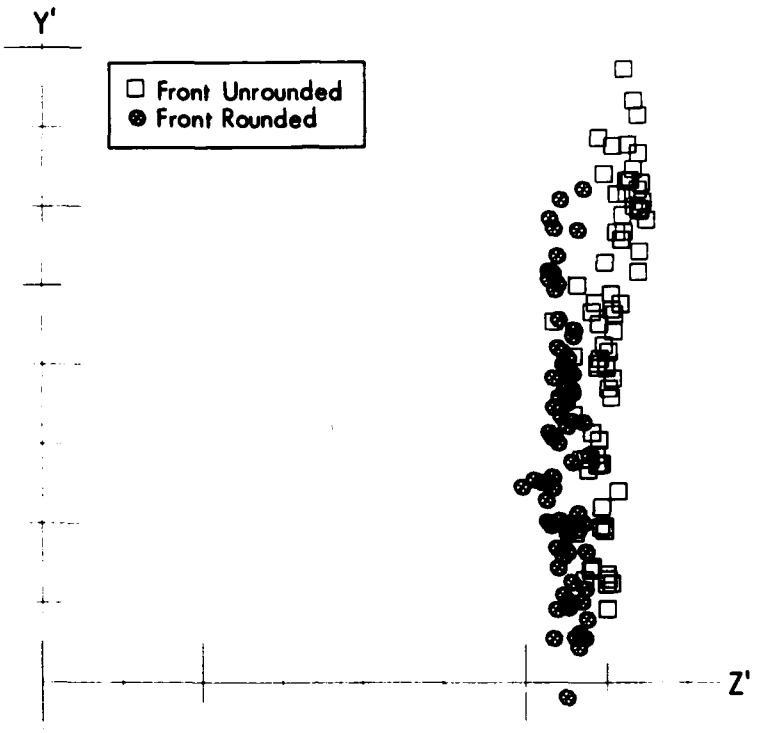


Fig. 8. Data points for the German front unrounded vowels ([i:, i, e:, e]) and the front rounded vowels ([y:, y, ø:, ø]) shown in side view using slab coordinates. The front unrounded vowels are represented by cubes, their rounded counterparts by spheres. In this orientation, the x' axis is perpendicular to the y'z' plane. x', y', and z' axes are in 0.1 log units and the point of origin is (0, 0, 0). The front of the space is to the right, with the data points for the rounded vowels falling behind those for the unrounded vowels.

with an APP of 0.58. While the percent correct classification of the x, y, and z variables does not reach the level of the hand-drawn target zones, this may be due to the fact that tokens are assigned to groups on the basis of group means. For target zones that envelope other zones, the group means will be very similar. In this sense, then, linear discriminant analysis would not be able to differentiate points falling, for example, in the target zones for German [a] and [a:] (see Figure 7), since these zones would have similar means. Nevertheless, the present analysis shows that vowel classification on the basis of x, y, and z reaches a level comparable to that on the basis of either F0, F1, F2, or F0, F1, F2, and F3. For another approach to the comparison of the APT scheme to other normalization schemes (e.g., Koenig, mel, and Bark), see Miller (1989).

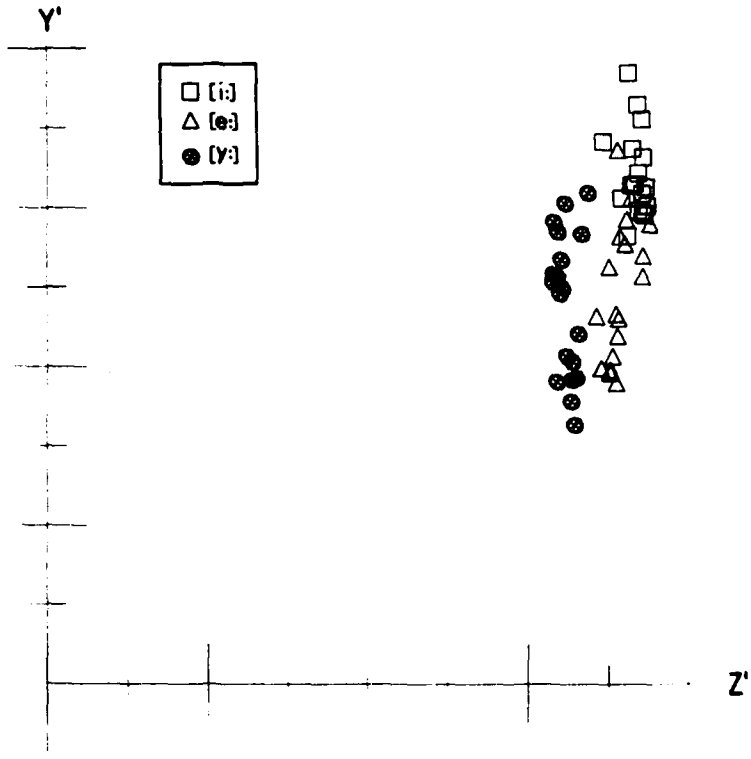


Fig. 9. Data points for the German unrounded vowels [i:] and [e:] along with the data points for the German rounded vowel [y:] shown in side view using slab coordinates. See Figure 8 for axis labels and units.

DISCUSSION

The present study attempts to characterize the vowels of Greek and German by converting F0, F1, F2, and F3 values into log frequency ratios and plotting them in the APS. Representing the relatively simple vowel inventory of Greek in the auditory-perceptual space enabled us to describe the five Greek vowels with 100% accuracy. The data points for the five Greek vowels are enclosed by non-contiguous target zones. Representing the much more complex German vowel system in the three-dimensional space enabled us to describe the vowels with 94% accuracy. The data points for the 14 German vowels are enclosed by adjacent target zones, with little or no unclaimed space between them. This representation of the German vowels yielded good separation along the front-back, rounded-unrounded, and tense-lax dimensions.

In his spectrographic study of the German vowels, Jørgensen (1969) pointed out that instead of representing the vowels in an F1 by F2 space, it might be advantageous to take F3 into account, especially for languages with both unrounded and rounded front vowels. Since, according to Jørgensen, three-dimensional displays are usually quite

hard to grasp, Jørgensen plotted the vowels in a two-dimensional space, using an effective F2' (Korlen and Malmberg, 1960), a measure that takes F3 into account. However, when Jørgensen plotted the German vowels in an F1 by F2' space, it did not enhance the separation of the unrounded and rounded front vowels at all, so that Jørgensen used an F1 by F2 space for the remainder of the paper (Jørgensen, 1969). Although Jørgensen did not quantify the amount of overlap in his two-dimensional space, it does seem that in the present analysis taking a third dimension (F3) into account as an independent dimension, rather than a derived dimension, substantially reduces the amount of overlap, particularly that between front unrounded and rounded vowels. Furthermore, the introduction of a low-frequency reference (SR) for speaker normalization and for the disambiguation of certain vowels may also have improved the classification of the 14 German vowels in the present study.

In addition, the three-dimensional space used in the auditory perceptual theory affords a way of comparing vowel systems across languages which is not subject to any of the objections brought forth by Disner (1980; 1986). Disner (1986) argues against normalization schemes which use mean formant frequencies as the correction factor when comparing vowel systems from different languages. She states that, when using overall formant frequency means for a language like French (or German) with front rounded vowels, the normalization procedure is likely to overnormalize the data, since there are more front vowels than back vowels. However, the auditory-perceptual theory does not use overall formant means as the correction factor for normalization. Thus, the vowels of each language can be mapped onto the same space, bounded by anatomical and physiological constraints, independent of vowel density in any area of the space.

It is possible, then, to represent and compare the vowels of all languages in a vowel space that is defined independently of vowel inventory. For example, the /i/ vowel of one language can be compared with the /i/ vowel of another language, regardless of the density of vowels in the language. It does seem to be the case that those vowel phonemes that the three languages discussed in this paper (Greek, German, and American English) have in common occupy similar locations in APS. Figure 10 shows a front view of APS with the three vowels that Greek, German, and American English have in common ([i, a, u]). For each language, each vowel is represented by a single point which is based on the average of the x', y', and z' values for all tokens of a particular vowel. The x', y', and z' values for American English [i, a, u] were taken from the database described in Miller (1989). As can be seen, these common vowels occupy similar regions in APS. However, it seems that the locations of these common vowels do, in fact, vary as a function of the phoneme inventory of each individual language, with the Greek vowels tending to be more centrally located and the German and American English vowels more peripherally located.

In general, it would seem advantageous for a given language to have vowels that are maximally distinct acoustically (see, for example, Liljencrants and Lindblom, 1972; Stevens, 1972; Lindblom, 1986) for reasons of communicative efficiency. Greek provides an example with its five vowels being quite far from each other in APS. Interestingly, five-vowel inventories similar to that of Greek are much more frequent than any other type of vowel inventory (Crothers, 1978; Maddieson, 1984). The question remains, then,

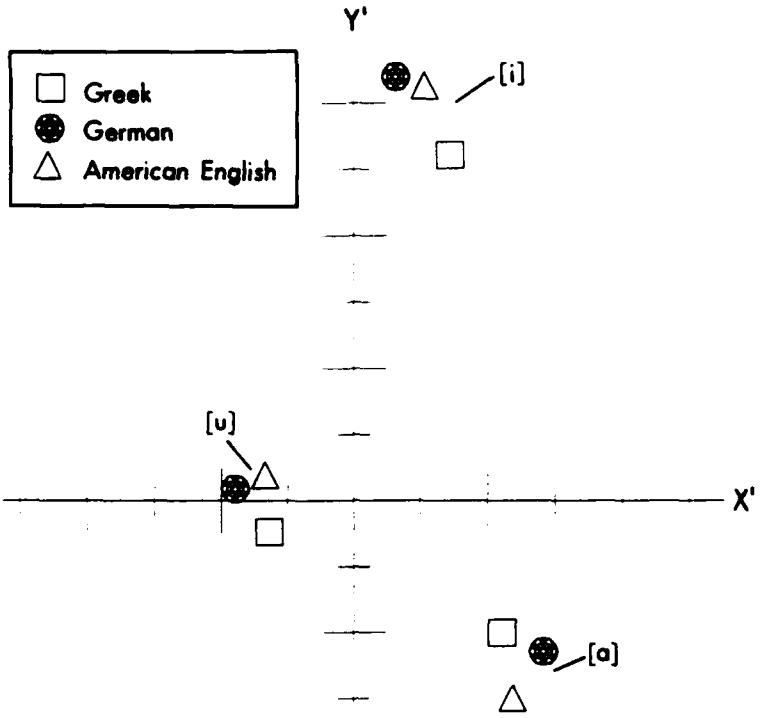


Fig. 10. Data points for three common vowels of Greek, German, and American English ([i, u, a]) shown in front view using slab coordinates. Each data point is the average of the x', y', and z' values of all tokens for each vowel in each language. See Figure 2 for axis labels and units.

how the vowels of languages with larger vowel inventories are organized in this same vowel space. The vowel spaces for German and American English are much more dense. It seems that the larger the vowel inventory, the more peripheral the location of the extreme vowels (in terms of x' and y'), relative to vowels of languages with smaller inventories.

As shown in Figure 10, this general trend holds for the vowels [i] and [u] of German, a language with 15 monophthongal vowels, as compared to those of American English, a language with 9 monophthongal vowels, which, in turn, are relatively more extreme than those of Greek, a language with 5 vowels. The exception is [a], which is most extremely located for American English. Of course, these preliminary findings will have to be replicated with a much larger sample.

In conclusion, the spectral characteristics of the five vowels in Modern Greek and 14 vowels in German were examined. For each vowel token, frequency values of F0, F1, F2, and F3 were obtained. A transformation was used to convert these measurements into log frequency ratios which were then plotted as points in the APS. Finally, target zones

were drawn around points representing phonemically identical vowels. These target zones could correctly categorize the present Greek and German corpora with 100% and 94% accuracy, respectively. Using this preliminary set of data, the present approach appears to normalize for a variety of inter- and intra-speaker variables and allow a unique characterization of vowel sounds in different languages, as well as a way to compare vowels across languages. In this manner, such an approach may provide insights not only into the language-specific organization of vowel systems, but also into cross-language comparisons.

(Received March 6, 1989; accepted October 6, 1989)

REFERENCES

ASSMANN, P.F., NEAREY, T.M., and HOGAN, J.T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America*, **71**, 975–989.

CROTHERS, J. (1978). Typology and universals of vowel systems. In J.H. Greenberg, C.A. Ferguson, and E.A. Moravcsik (eds.), *Universals of Human Language, Vol. 2: Phonology* (pp. 93–152). Stanford: Stanford University Press.

DISNER, S.F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, **67**, 253–261.

DISNER, S.F. (1986). On describing vowel quality. In J.J. Ohala and J.J. Jaeger (eds.), *Experimental Phonology* (pp. 69–79). New York: Academic Press.

FANT, G. (1973). *Speech Sounds and Features*. Cambridge: MIT Press.

HOUSEHOLDER, F.W., KAZAZIS, K., and KOUTSOUDAS, A. (1964). *Reference Grammar of Literary Dhimotiki*. IJAL 30/2 Part II, Publications of the Indiana University Research Center in Anthropology, Folklore, and Linguistics, 31. Bloomington: Indiana University.

IIVONEN, A. (1987). A set of German stressed monophthongs analyzed by RTA, FFT, and LPC. In R. Channon and L. Shockey (eds.), *In Honor of Ilse Lehiste* (pp. 125–138). Dordrecht: Foris Publications.

JØRGENSEN, H. (1969). Die gespannten und ungespannten Vokale in der norddeutschen Hochsprache, mit einer spezifischen Untersuchung der Struktur ihrer Formantenfrequenzen. *Phonetica*, **19**, 217–245.

KORLEN, G., and MALMBERG, B. (1960). *Tysk Fonetik*. Lund: Gleerups.

LILJENCRANTS, J., and LINDBLOM, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, **48**, 839–862.

LINDBLOM, B. (1986). Phonetic universals in vowel systems. In J.J. Ohala and J.J. Jaeger (eds), *Experimental Phonology* (pp. 13–44). New York: Academic Press.

MADDIESON, I. (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.

MILLER, J.D. (1987a). Auditory-perceptual processing of speech waveforms. In W.A. Yost and C.S. Watson (eds.), *Auditory Processing of Complex Sounds* (pp. 257–266). Hillsdale: Erlbaum.

MILLER, J.D. (1987b). Classification of vowel production by means of perceptual target zones: A response to Ladefoged and Studdert-Kennedy. *Journal of the Acoustical Society of America*, **82**, Suppl. 1, S82.

MILLER, J.D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, **85**, 2114–2134.

MOULTON, W.G. (1962). *The Sounds of German and English*. Chicago: University of Chicago Press.

NEAREY, T.M. (1978). *Phonetic Feature Systems for Vowels*. Bloomington: Indiana University Linguistics Club.

PETERSON, G.E. (1952). The information bearing elements of speech. *Journal of the Acoustical Society of America*, **24**, 629–637.

PETERSON, G.E., and BARNEY, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175–184.

POLS, L.C.W. (1977). *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words.* Soesterberg, The Netherlands: Institute for Perception TNO.

SHEPARD, R.N. (1972). Psychological representation of speech sounds. In E.E. David and P.B. Denes (eds.), *Human Communication: A Unified View* (pp. 67–113). New York: McGraw Hill.

STEVENS, K.N. (1972). The quantal nature of speech. In E.E. David and P.B. Denes (eds.), *Human Communication: A Unified View* (pp. 51–66). New York: McGraw Hill.

SYRDAL, A.K. (1985). Aspects of a model of the auditory representation of American English vowels. *Speech Communication*, **4**, 121–135.

SYRDAL, A.K., and GOPAL, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086–1100.

WIESINGER, P. (1983). Die Einteilung der deutschen Dialekte. In W. Besch, U. Knoop, W. Putschke, and H.E. Wiegand (eds.), *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung, Vol. 2* (pp. 807–900). Berlin: Walter de Gruyter.

WRIGHT, J.T. (1986). The behavior of nasalized vowels in the perceptual vowel space. In J.J. Ohala and J.J. Jaeger (eds.), *Experimental Phonology* (pp. 45–67). New York: Academic Press.

APPENDIX

Tables A1 and A2 list the fundamental frequency and formant frequency values for each of the vowel tokens plotted in the APS. Each measurement reported represents the geometric mean of the relevant parameter over the total duration of the steady-state portion of the vocalic section transcribed as the intended vowel. To replicate the plots in the main paper, one needs to convert the F0 values into SR values using Equation 4 and then use Equations 1 through 3, or Equations 5 through 7, to derive x, y, z or x', y', and z' coordinates, respectively. Table A1 lists the values for the five Greek vowels and Table A2 for the 14 German vowels.

TABLE A1

		Vowel [i]				Vowel [e]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	192	317	2226	2550	182	475	1806	2247
	2	203	319	2337	2635	197	477	1821	2300
	3	216	311	2345	2603	206	484	1854	2276
	4	201	317	2088	2320	200	481	1818	2130

2.	1	169	311	2099	2374	172	469	1626	2056
	2	157	308	2093	2354	157	468	1657	1989
	3	197	311	2134	2362	193	510	1625	2043
	4	195	312	2145	2332	202	479	1614	2038
3.	1	128	310	1725	2188	124	468	1407	2425
	2	129	303	1729	2183	130	473	1408	2422
	3	130	304	1777	2075	123	469	1407	2418
	4	130	300	1716	2266	127	456	1414	2466
4.	1	135	310	2037	2549	131	471	1676	2322
	2	143	311	2045	2644	135	473	1703	2460
	3	124	301	2103	2655	125	472	1667	2273
	4	125	310	2042	2617	143	466	1749	2565

		Vowel [a]				Vowel [o]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	182	770	1300	2329	191	470	802	2433
	2	180	771	1368	2320	194	476	814	2428
	3	180	773	1414	2397	198	474	811	2492
	4	198	736	1351	2177	197	482	825	2484
2.	1	156	739	1138	2970	150	471	750	1857
	2	148	718	1171	3027	174	473	856	1853
	3	146	674	1146	3078	175	470	887	1979
	4	195	626	1175	2899	197	481	762	2086
3.	1	123	626	1125	2457	130	473	945	2519
	2	118	595	1188	2509	130	476	837	3239
	3	122	594	1151	2509	131	480	825	2733
	4	117	625	1157	2382	129	473	878	2676

4.	1	132	683	1242	2326	136	486	938	2198
	2	134	645	1242	2275	133	478	959	2224
	3	128	663	1248	2380	138	482	866	2303
	4	129	648	1245	2307	140	477	903	2351

Vowel [u]

Spkr.	Token	F0	F1	F2	F3
1.	1	215	421	831	2490
	2	212	421	728	2491
	3	225	350	616	2491
	4	215	384	609	2461
2.	1	168	312	855	2240
	2	165	312	771	2107
	3	209	342	749	2139
	4	216	328	730	2203
3.	1	139	313	907	2336
	2	133	312	1509	2302
	3	139	311	1000	2300
	4	131	312	1001	2295
4.	1	148	358	1004	2648
	2	148	315	1065	2471
	3	134	312	850	2361
	4	147	313	836	2338

TABLE A2

		Vowel [i:]				Vowel [i]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	198	315	2556	3306	184	463	2048	2633
	2	199	310	2672	3452	184	353	1798	2415
	3	188	307	2507	3372	185	329	1769	2227
	4	194	306	2607	3117	195	453	2030	2458
2.	1	214	302	2562	3479	177	459	2327	2787
	2	213	311	2526	3494	199	434	2297	2795
	3	205	317	2576	3447	206	443	2390	2855
	4	202	315	2625	3445	188	360	2138	2802
3.	1	249	304	2594	3400	269	465	1926	2937
	2	238	299	2652	3476	244	465	1915	2801
	3	258	308	2504	3301	260	472	1927	2814
	4	249	308	2646	3515	244	468	1939	3011

		Vowel [e:]				Vowel [e]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	176	347	2463	3262	177	432	1971	2546
	2	202	388	2379	2997	180	624	1963	2525
	3	181	315	2312	3277	184	629	2034	2607
	4	177	311	2499	3377	188	617	2006	2638
2.	1	204	430	2539	3061	204	648	2062	2930
	2	190	344	2573	3025	200	642	2120	2970
	3	206	441	2689	3043	196	680	2065	2932
	4	189	324	2539	3099	195	629	2032	2968
3.	1	254	470	2483	2969	229	474	2077	2933
	2	258	465	2469	3113	224	596	2039	2998
	3	255	468	2509	3106	237	574	2029	3002
	4	241	469	2468	3160	249	488	2045	3028

		Vowel [y:]				Vowel [y]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	195	340	1727	2296	192	476	1686	2296
	2	196	333	1583	2199	191	333	1783	2417
	3	188	310	1509	2352	184	467	1700	2568
	4	193	309	1717	2453	189	498	1570	2149
2.	1	227	304	1535	2500	219	459	1565	2502
	2	220	306	1551	2486	221	456	1506	2478
	3	222	307	1558	2496	220	470	1666	2575
	4	220	308	1353	2517	215	468	1653	2649
3.	1	243	312	1872	2464	212	398	1569	2230
	2	264	308	1879	2427	292	464	1614	2150
	3	265	309	1875	2409	279	464	1703	2249
	4	257	310	1842	2392	263	467	1562	2392

		Vowel [ø:]				Vowel [ø]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	189	319	1579	2335	186	637	1747	2429
	2	194	333	1559	2338	184	557	1888	2566
	3	184	314	1436	2306	188	616	1846	2551
	4	180	313	1429	2215	181	589	1797	2392
2.	1	213	460	1477	2315	199	628	1725	2653
	2	206	445	1411	2365	216	627	1708	2646
	3	213	459	1406	2327	205	629	1662	2599
	4	204	437	1412	2402	201	625	1708	2609
3.	1	266	427	1576	2341	274	611	1712	2469
	2	269	466	1701	2292	232	659	1712	2337
	3	254	465	1629	2309	234	666	1560	2532
	4	246	467	1712	2344	232	505	1711	2341

		Vowel [u:]				Vowel [u]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	185	317	840	1935	182	345	853	2227
	2	179	306	1189	1954	183	347	742	2305
	3	182	316	956	2047	187	333	758	1932
	4	180	315	843	1883	177	318	761	2239
2.	1	207	329	1026	2567	205	458	801	2770
	2	214	305	808	2641	208	467	800	2778
	3	198	317	632	2600	209	463	798	2730
	4	202	357	708	2488	203	456	755	2804
3.	1	274	312	713	2552	265	491	739	2552
	2	253	314	589	2532	248	478	931	2529
	3	250	313	568	2568	237	469	932	2451
	4	245	315	598	2496	242	468	894	2418

		Vowel [o:]				Vowel [o]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	182	332	756	2917	168	541	1391	2491
	2	171	317	598	2610	167	528	1262	2483
	3	171	316	600	2442	178	487	1251	2324
	4	166	321	834	2567	172	611	1332	2435
2.	1	189	371	758	2894	182	697	1104	2667
	2	–	–	–	–	187	737	1191	2678
	3	180	325	626	2820	181	668	1129	2653
	4	189	354	721	2834	175	658	1123	2735
3.	1	271	483	718	2406	259	809	1244	2510
	2	236	468	964	2437	228	637	1133	2558
	3	242	472	1017	2466	217	636	1109	2522
	4	247	469	917	2485	238	642	1184	2413

		Vowel [a:]				Vowel [a]			
Spkr.	Token	F0	F1	F2	F3	F0	F1	F2	F3
1.	1	174	754	1505	2464	177	676	1708	2534
	2	176	800	1577	2602	180	723	1851	2537
	3	180	777	1409	2592	169	654	1610	2542
	4	172	800	1350	2693	189	727	1531	2500
2.	1	190	796	1259	2703	204	816	1361	2752
	2	184	802	1258	2653	199	813	1405	2671
	3	175	805	1289	2732	191	795	1491	2791
	4	192	819	1370	2739	203	974	1439	2675
3.	1	231	957	1288	2514	261	1084	1485	2670
	2	228	849	1167	2527	252	1003	1420	2524
	3	249	1188	1316	2662	240	967	1405	2522
	4	235	982	1315	2604	249	995	1408	2566

Approved for public release;
distribution unlimited.

955
AFOSR Grant G-AFOSR-86-0335
Final Technical Report
Appendix

Plosive/fricative distinction: The voiceless case

LaDeana F. Weigelt, Steven J. Sadoff, and James D. Miller
Central Institute for the Deaf, 818 South Euclid Avenue, St. Louis, Missouri 63110

(Received 31 August 1989; accepted for publication 22 January 1990)

Using only three measures of the waveform, the zero-crossing rate, the logarithm of the root-mean-square (rms) energy, and the derivative of the log rms energy with respect to time [termed rate of rise (ROR)], voiceless plosives (including affricates) can be distinguished from voiceless fricatives in word-initial, medial, and final positions. Peaks in the ROR contour are considered for significance to the plosive/fricative distinction by examining the log rms energy and zero-crossing rate. Then, the magnitude of the first significant peak in the ROR contour is used as the primary classifier. The algorithm was tested on 1364 tokens (720 word-initial tokens produced by four female and four male speakers; 360 word-medial tokens produced by two males and two females; 320 word-final tokens produced by two males and two females). Data from two male and two female speakers (360 word-initial tokens) were used as a training set, and the remaining data were used as a test set. The overall rate of correct classification was 96.8%. Implications of this result are discussed.

PACS numbers: 43.72.Ar, 43.70.Fq, 43.72.Ne

INTRODUCTION

The purpose of this paper is to describe a method for automatically distinguishing voiceless plosives (including affricates) from voiceless fricatives using the rate of rise (ROR) of the logarithm of the root-mean-square (rms) energy of the acoustic waveform. While investigators have postulated several acoustic cues for this distinction, no reliably detectable cue has been systematically shown to accurately distinguish these classes of sounds.

Possible applications of an algorithm to distinguish voiceless plosives from voiceless fricatives include classification of speech segments into broad phonemic categories as employed by many automatic speech recognition systems (Reddy, 1966; Weinstein *et al.*, 1975; Schwartz and Makhoul, 1975; Paliwal and Rao, 1977; Regel, 1982; Paliwal and Rao, 1982). For example, Reddy groups segments into four nonmutually exclusive subsets: stop-like sounds, fricative-like sounds, nasal-liquid-like sounds, and vowel-like sounds. Groupings like these have also found application in lip-reading aids such as the wearable eyeglass speech-reading aid described by Upton (1968).

Since the gross spectral shapes of plosives and fricatives are similar, several other cues to distinguish between these classes of voiceless consonants have been suggested, all based on properties of the amplitude envelope. These cues include consonant duration (Gerstman, 1957; Weinstein *et al.*, 1975), burst amplitude (Repp, 1984; Dorman *et al.*, 1980), silent closure duration (Repp *et al.*, 1978; van Heuven, 1987; Repp, 1984), steady time defined as the duration of the portion of the consonant where there is minimal spectral change (Gerstman, 1957), and rise time (Gerstman, 1957; Cutting and Rosner, 1974, 1976; Pickett, 1980; Stevens, 1980). In this paper, a variable related to rise time, the rate of rise of the log rms energy of the waveform, is the primary measure used for the distinction.

The ROR of log rms energy has the potential to be particularly useful for the plosive/fricative distinction (voiceless affricates are treated as plosives because they begin with a plosive burst) because it takes advantage of the differences in onset characteristics of plosives and fricatives. Plosives and affricates, whose properties can be found in the literature (Halle *et al.*, 1957; Pickett, 1980; Howell and Rosen, 1983; Repp, 1984), are seen to have an abrupt onset and therefore a high ROR value. Although this abrupt amplitude change has been noted elsewhere (for example, Stevens, 1980), most earlier observations were qualitative in nature. Fricatives, whose general characteristics are discussed elsewhere (Strevens, 1960; Heinz and Stevens, 1961; Shadle, 1985), are seen at onset to have a gradual rise in log rms energy and therefore a relatively low ROR value. Thus an ROR threshold can be set such that waveforms which have a peak in their ROR contour (correlated with consonant onset) with a value above the threshold are labeled as plosives, and those with a peak ROR value below the threshold are labeled as fricatives.

The problem of distinguishing voiceless plosives from voiceless fricatives, then, reduces to determining which ROR peaks are correlated with consonant onset. In most cases, the highest peak in the ROR contour is the peak in question. However, large ROR peaks resulting from nonspeech sounds (i.e., lip smacks or saliva pops) preceding the syllable can be confused with consonant onset. For the application discussed here, these spurious peaks are rejected using methods similar to those commonly used for signal endpoint detection, i.e., setting thresholds for relative energy magnitude, energy duration, and zero-crossing rate of the signal (Rabiner and Sambur, 1975; Lamel *et al.*, 1981).

I. METHODS

A. Speakers and tokens

Eight native Americans, four male and four female, with no known history of either speech or hearing disorders served as speakers. Of these speakers, only two had a formal

 0001-4966/90/062729-09$00.80 © 1990 Acoustical Society of America

course in phonetics, one had some exposure to phonetic transcription, and five had no exposure to phonetics. The word lists were of the form VCV or CVC, where the consonant of interest was a voiceless consonant. Separate recordings were made for each word position. The word lists are described here: word-initial position (90 CVC tokens), 9 consonants (3 plosives /p,t,k/, 5 fricatives /s,ʃ,f,θ,h/, and 1 affricate /tʃ/) × 10 vowels /i,ɪ,ɛ,æ,a,ʌ,ɔ,ʊ,u,ɝ/; word-medial position (81 VCV tokens), 3 vowels /i,a,u/ × 9 consonants /p,t,k,s,ʃ,f,θ,h,tʃ/ × 3 vowels /i,a,u/; and word-final position (80 CVC tokens), 10 vowels × 8 consonants /p,t,k,s,ʃ,f,θ,tʃ/. For word-final position only eight consonants were used since /h/'s do not occur in word-final position in American English. Also, because the affricate /ts/ is not phonemic in American English, it was not included in the data set.

B. Recordings

The recordings were made in an anechoic chamber using a low-noise microphone/preamplifier combination (Bruel & Kjaer 4179/2660). The microphone was placed at a height equal to and ½ m in front of the subject's mouth (zero degrees angle of incidence). In this setup, turbulent noise generated by the breath flow over the microphone was not a factor. Conversational speech levels were used, i.e., 60–65 dBA to the microphone. The microphone output was channeled directly into a Sony PCM-501ES digital audio recorder (16-bit mode) with a JVC 720 VCR serving as the storage medium.

Prior to the recording, speakers familiarized themselves with the list of randomized words. For the VCVs, talkers were instructed to place the emphasis on the second syllable.

Speakers began reading tokens while an appropriate recording level was selected. When that recording level had been set, a calibration tone was recorded for that particular subject. The calibration tone consists of a 1-kHz sine wave generated by a Hewlett–Packard 3325A synthesizer and maintained at a constant output level of 69.5 mV at the input of the Sony digital audio recorder (equivalent to 70 dB SPL at the microphone).

C. Analysis

The recordings were played back from the Sony recorder in analog form and redigitized with proper antialiasing. The recordings were digitized using a Micro Technology Unlimited Digisound-16 analog-to-digital and digital-to-analog converter and a custom hardware interface developed in-house. Digitization was performed at 20 kHz with 16-bit precision and edited to include 500 ms per token. For the word-initial consonants, this 500 ms included time before the utterance, the initial plosive/affricate or fricative with its corresponding transition region, and at least 50 ms of the following vowel. For VCVs, the entire initial vowel and medial consonant as well as the onset of the final vowel was digitized. In the word-final position case, the 500 ms digitized included a portion of the medial vowel and the entire final consonant. The files were each notch filtered at 60 Hz to remove ac noise and stored on a Micro VAX II.

Productions from all eight speakers were used for the word-initial study. Two male and two female speakers from the word-initial study were used for the word-medial and word-final studies. In the word-initial case, the productions were divided into two sets: All utterances from two males and two females were chosen as the training set with the 360 tokens recorded from the remaining four speakers serving as a test set. For the word-medial and final positions, the algorithm developed and trained on the word-initial training set was simply applied; i.e., there was no retraining.

II. ALGORITHM

The algorithm was developed using a combination of measures that could be easily calculated and efficiently obtained. Throughout this paper, indices within parentheses refer to sample number, while subscripts refer to frame number. First, the input waveform is preemphasized using the first-order difference equation, $y(i) = x(i) - 0.98x(i-1)$, where $x(i)$ is the ith input sample. This one–zero filter, $H(z) = 1 - 0.98z^{-1}$, acting as a differentiator, emphasizes the high-frequency components where most of the acoustic cues for voiceless consonants reside. Paliwal (1984), in his work using minimum-distance classifiers for phoneme recognition, suggests that preemphasis should be used for consonant discrimination.

The short-term log rms energy E_n is defined as

$$E_n = 10 \log\left[\left(\frac{1}{N}\right) \sum_{i=cn}^{cn+N-1} [y(i)w(i-cn)]^2\right], \qquad (1)$$

where $w(i)$ is the standard finite-duration Hamming window (Oppenheim and Schafer, 1975, p. 242):

$$w(i) = 0.54 - 0.46 \cos[2\pi i/(N-1)], \quad 0 \leqslant i < N. \qquad (2)$$

For the Hamming window a value of $N = 480$ samples (24-ms frames) is used. This window is moved in $c = 20$ sample (1-ms) steps.

The primary measure used is the rate of rise of log rms energy, which is defined as

$$\mathrm{ROR}_n = (E_n - E_{n-1})/(\Delta t), \qquad (3)$$

where Δt is the separation (in seconds) of the energy measurements. In our work, Δt = (20 samples)/(20 000 samples/s) = 0.001 s.

The zero-crossing rate is calculated from the raw input waveform, before preemphasis. It has been shown (Paliwal *et al.*, 1983) that for speech recognition the zero-crossing rate of the raw signal is better than the zero-crossing rate of the preemphasized version. Because of the recording conditions for this study (i.e., an anechoic chamber with high-quality recordings), the effects of background noise and dc bias are minimal. Hence, the calculation of the number of times the input signal crosses the axis (zero-crossing rate) is an effective parameter in detecting unvoiced speech. Clearly, applying a Hamming window does not alter the number of zero crossings since the sign of the signal is not changed.

Peaks are located in the ROR function and are sorted in descending order by amplitude. Only peaks with positive

ROR values need to be considered since peaks with negative ROR values correspond to offset rather than onset of the consonant. Each of these peaks is considered in order (largest peak first) as a potential candidate. If a candidate peak passes each of four criteria to be described below and justified in Sec. IV A, then a significant peak is found. If a peak fails any of the criteria, then it is discarded as being nonsignificant. The rate-of-rise value of the first significant peak is used as the peak ROR value for that token. Sounds with peak ROR values above a predetermined cutoff are classified as plosives, while sounds with values below the cutoff are labeled as fricatives. If no significant peaks are found, the segment is labeled as a fricative.

In most cases, the largest peak in the ROR contour corresponds to the onset of the plosive or fricative and thus is the "significant" peak. However, in some cases, ROR peaks corresponding to nonspeech sounds or vowel onsets can also have large values and thus can be confused with consonant onset. By examining training set data and visually choosing criteria and then testing these criteria, we were able to eliminate these spurious peaks. While the exact values specified in the criteria may not be crucial and perhaps other combinations of measures may be equally effective, the criteria described below have proven to give excellent results.

For a candidate ROR peak to be labeled as significant, then, the following four conditions must be met. First, for the 49-ms period[1] following the peak, the value of E_n must never fall below the value of E_n at the peak. Second, the maximum value of E_n for the following 49 ms must be at least 12 dB above the value of E_n at the peak. Third, the maximum zero-crossing rate over the 49-ms period after the peak must be greater than 2000 zero crossings per second. Finally, the zero-crossing rate, exactly 49 ms after the peak, must be no more than 100 crossings per second below the zero-crossing rate 20 ms before the peak.

III. RESULTS

Examination of algorithm performance on both the training and test sets suggests the validity of using log rms energy, ROR, and zero-crossing rate for the plosive/fricative distinction. The performance of the algorithm can be better appreciated by looking at easy to label examples of both a typical plosive and fricative, as well as more difficult to label examples. In Fig. 1, a typical plosive (/k/) can be seen with its characteristic abrupt increase in log rms energy at plosive onset and the corresponding sharp peak in the ROR contour. The peak labeled with the arrow is the "significant" peak and has an ROR value above the threshold; thus it is correctly labeled as a plosive. In Fig. 2, the typical fricative (/ʃ/) is seen to have a gradual rise in log rms energy, rather than the abrupt rise seen for the plosive, and low peak values in the ROR contour. Again, the peak identified by the arrow was chosen as the "significant" peak, and its ROR value was below the ROR threshold; thus it is labeled as a fricative.

Two examples which would be difficult to label by visual inspection of the waveforms can be seen in Figs. 3 and 4. The plosive (/p/), seen in Fig. 3, has a gradually rising energy contour similar to that of a fricative. However, its ROR value is above the ROR threshold (see Sec. III A) and thus is labeled correctly as a plosive. The fricative seen in Fig. 4 (/s/) has an energy contour which looks similar to that seen for the plosive in Fig. 3. However, its ROR value for the significant peak falls below the threshold value and is thus correctly labeled as a fricative. Also in Fig. 4, several nonspeech energy pulses prior to the fricative onset can be seen, which were correctly rejected using the significant peak criteria. The ROR peaks correlated with the first four lobes in the energy contour are rejected as nonspeech sounds because, during the 49-ms period following the peaks, the value

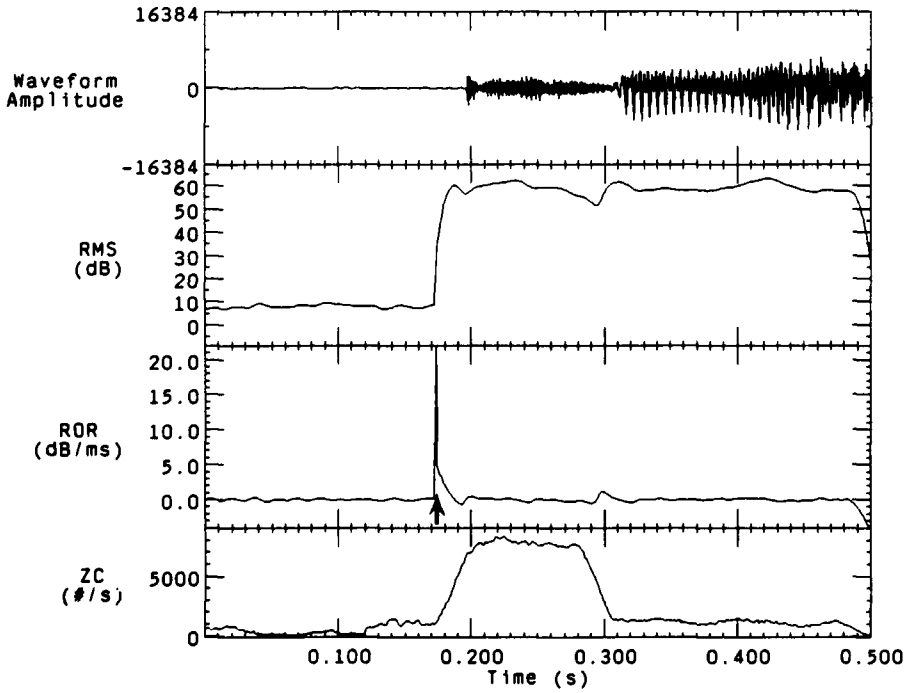


FIG. 1. An example of a plosive, the /k/ from the token /kil/, typical of the plosives seen in this study and correctly identified by the algorithm as a plosive. The waveform, log rms energy, ROR of log rms energy, and zero-crossing rate are shown. Notice the rapid ROR of log rms energy at the plosive onset signaled by the peak in the ROR plot. The peak denoted by the arrow is the "significant" peak. The apparent discrepancy between the position of the plosive onset in the waveform and the peak in the ROR contour results from the fact that the window used for the calculation of the log rms energy is defined in terms of its leftmost point.

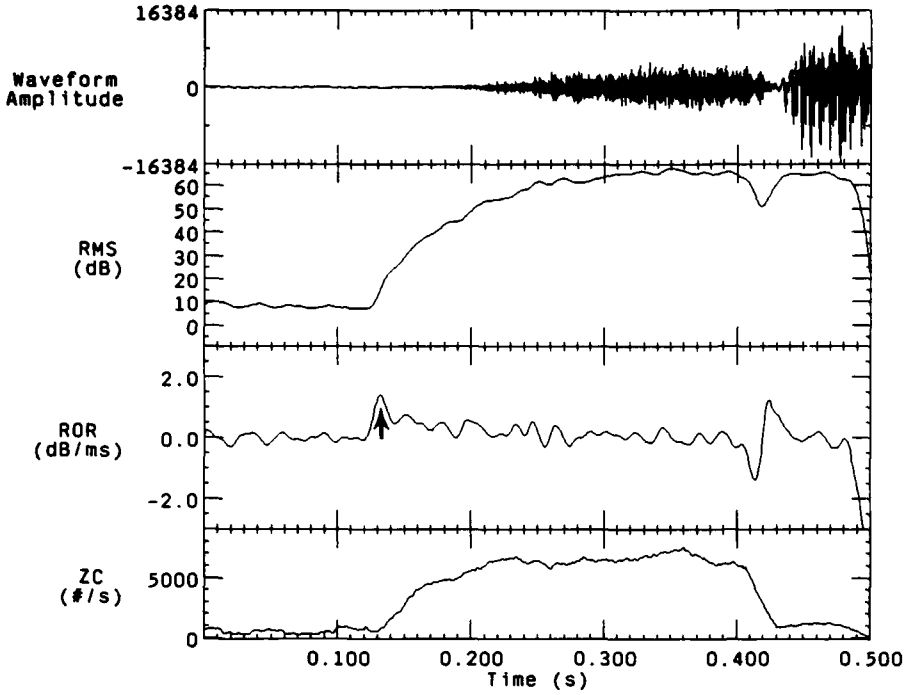


FIG. 2. An example of a typical fricative, the /ʃ/ from the token /ʃIp/, its waveform, log rms energy, ROR, and zero-crossing rate, correctly identified by the algorithm as a fricative. Notice the gradual rise of the energy contour correlated with a small peak (see arrow) in the ROR plot. Note that the peak in the ROR plot corresponding to the vowel onset was correctly rejected; it was not a "significant" peak.

of E_n falls below the value of E_n at the peak and the maximum zero-crossing rate is not greater than 2000 zero crossings per second. Thus cases which could be difficult to label by hand can be correctly labeled using this algorithm.

A. ROR threshold

The ROR threshold for the plosive/fricative distinction was determined using the word-initial training set of 360 tokens. The percentile plot of peak ROR values for the plosives, fricatives, and affricate (Fig. 5) provides a visual method of obtaining an ROR cutoff for the plosive/fricative distinction. Several possible ROR threshold values could be chosen, thereby trading misclassifications of plosives with those of fricatives. The highest percentage of correct classifications for the training set is obtained when the ROR threshold value is set between 2.215 and 2.265 dB/ms. Since it has been found that the plosives /p,t,k/ appear equally as often in conversational speech as do the fricatives /s,ʃ,f,θ,h/ (Carterette and Jones, 1974), the ROR cutoff was chosen at the midway point of this ambiguity region, 2.24 dB/ms.

In Fig. 6, peak ROR values separated by phoneme can

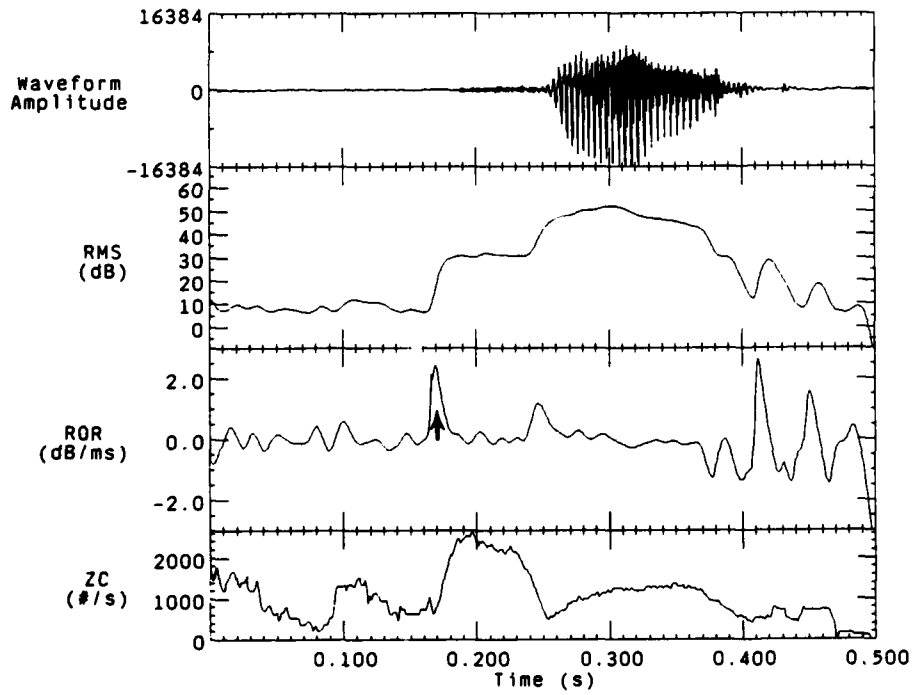


FIG. 3. An example of a plosive, the initial plosive in /pUt/, whose log rms energy looks similar to that of a fricative. With the ROR threshold set at 2.24 dB/ms, however, the ROR value of the significant peak allows it to be correctly labeled as a plosive.

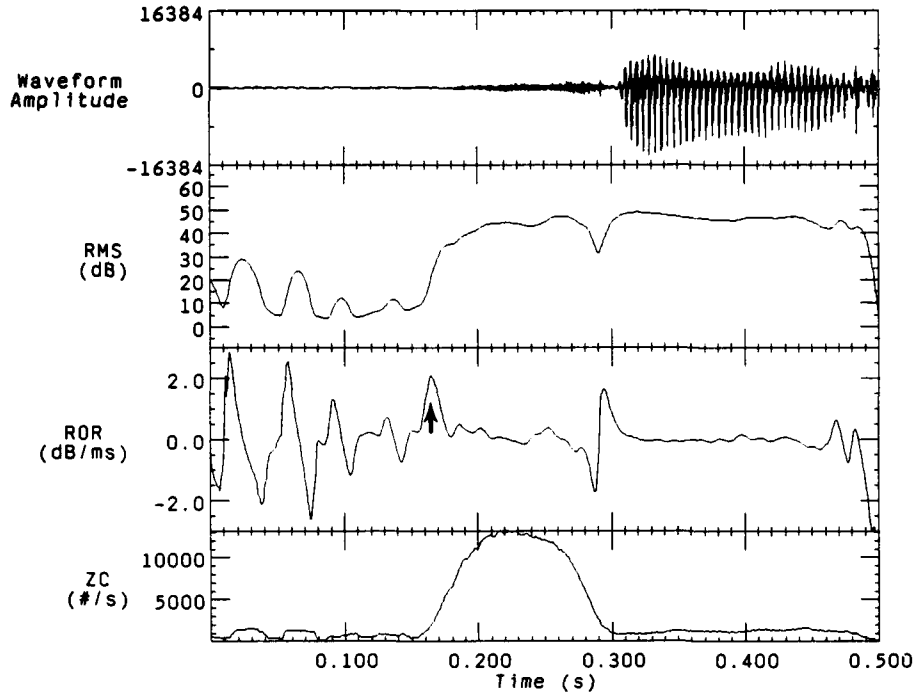


FIG. 4. An example of a fricative, the /s/ from the token /s⊃t/, whose energy contour is similar to that of a plosive. With the ROR threshold set at 2.24 dB/ms, however, it is correctly labeled as a fricative. Notice also that several peaks early in the waveform, corresponding to nonspeech sounds, were correctly rejected; that is, they were not labeled as significant peaks.

be seen. These data indicate that not only can ROR be used to separate plosives and fricatives, but it may, in the case of the plosives, also be used as a contributing cue to identify place of articulation. Among the plosives, the highest ROR values are observed for /t/ bursts and intermediate values are observed for the /k/ bursts, while the lowest values are observed for the /p/ bursts. While it is difficult to separate the fricatives on the basis of ROR, it is noted that /s/ and /ʃ/, on the average, have higher peak ROR values than the

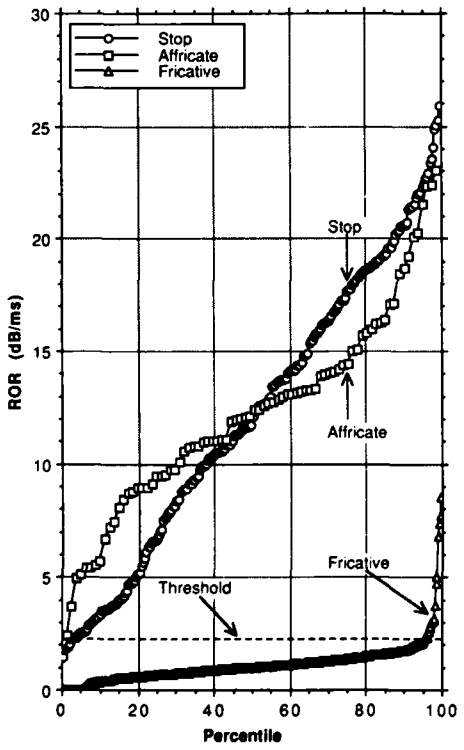


FIG. 5. A plot of ROR values in dB/ms versus percentile ranks separated by manner of articulation for the 720 word-initial tokens. It can be seen that plosives and affricates have higher ROR values than do fricatives. Percent correct scores resulting from each choice of an ROR threshold can also be obtained.

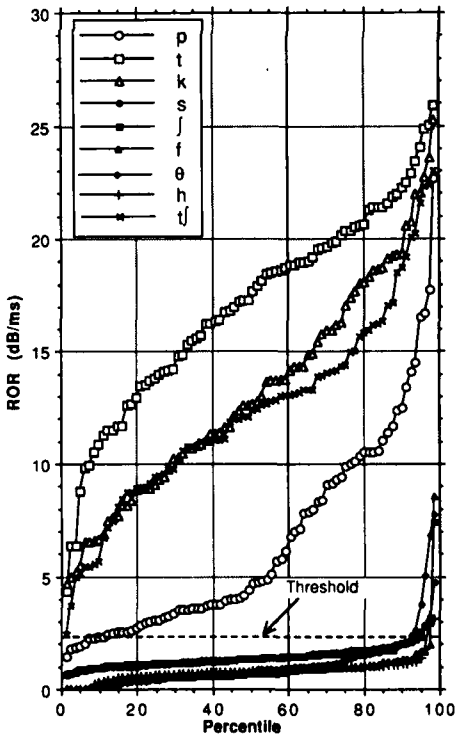


FIG. 6. A plot of ROR values (dB/ms) versus percentile ranks for 720 word-initial tokens separated by place of articulation. It can be seen that of the voiceless plosives, /t/ has the highest ROR values, /k/ has the intermediate values, and /p/ has the lowest. Fricative ROR values are so similar that they are impossible to distinguish on the graph. However, it can be noted that /ʃ/ has the highest ROR values of the group. Overlap in ROR values, then, primarily occurs between /p/ and /ʃ/.

others. Some higher ROR values for fricatives are due to nonspeech sounds occurring simultaneously with fricative onset, which sometimes causes a fricative to be mislabeled as a plosive.

B. Results for each consonant position

Across all word positions (including testing and training sets), the algorithm has a 96.8% rate of success (Table I). The success rate on the word-initial sets (testing and training) also reached 96.8%. In the training set (360 tokens), only 8 (2.2%) were mislabeled. Of those errors, 7 were fricatives mislabeled as plosives, and in each case the onset of the fricative occurred simultaneously with a nonspeech sound. In the test set (360 tokens), 15 (4.2%) were mislabeled. Of these, 11 were fricatives and 4 were plosives.

There are some possible explanations for the differences in percent success between the word-initial training and test sets. First, the algorithm, especially the "significant" peak criteria, was developed by examining training set data. Thus performance should be better on that set. Second, the speakers used for the training set were experienced subjects familiar with the recording environment. The speakers who produced the test set data, however, were without such experience and seemed more nervous and uncomfortable in the recording situation. Therefore, they may have been less precise in their productions and may have produced more extraneous sounds at the time of recording. In the light of this possibility, a trained phonetician was asked to listen to and transcribe the initial voiceless consonant of the 720 word-initial tokens presented in random order. The 360 token training set was transcribed with 100% accuracy, suggesting that any mislabeled tokens were actually algorithm error. However, in the test set, five of the fricatives were transcribed as plosives. These five fricatives were also labeled as plosives by the algorithm. These five tokens, then, could be removed from the test set (because of the contradiction in labeling) or counted as correct. In either case the success rate of the algorithm rises to 97.2% for the test set. While these data could be "thrown away," establishing a consistent objective criterion for discarding data is difficult, if not impossible. Therefore, all original data were retained.

Performance on word-medial and word-final positions was also less accurate than that on the word-initial training set. In word-medial position, ten errors were mislabeled /p/'s (in addition there was one mislabeled affricate), each of whose peak ROR value was between 1.50 and 2.16 dB/ms. In the word-final case, there were eight /p/'s and one affricate that were mislabeled, again with low peak RORs. However, the percentages correct for word-medial and final positions of the consonants of 96.6% and 97.2% can be increased to 99.1% and 98.8%, respectively, if the ROR threshold is adjusted to 1.62 dB/ms rather than 2.24 dB/ms. There are two possible explanations for the increase in percent correct brought about by lowering the ROR threshold. One explanation is the fact that adjusting the ROR threshold specifically for individual data sets increases percentages by better predicting the error term. However, a more likely explanation is that plosives in word-medial (with stress placed on the syllable containing the medial consonant) and final position have, in general, a lower average ROR value than those in word-initial position (especially when the word-initial token is spoken as an isolated syllable where more time is available for a buildup of pressure behind the point of closure). Support for the latter hypothesis is provided in Fig. 7 where it can be seen that the average ROR value for initial position plosives is 12.2 dB/ms, whereas for medial and final positions it is 9.8 dB/ms. This implies that in a future implementation of this algorithm it may be useful to lower the ROR threshold when the plosive (or affricate) is in word-medial or final position.

C. ROR versus manner

The results across all word positions were examined using a three-factor ANOVA (ROR versus manner, sex, and vowel). The differences in mean ROR values for the different manners of articulation are shown in Fig. 7 and are significant at $\alpha = 0.0001$. In particular, using Tukey's procedure, it was found that the differences between the means of the plosives (mean = 11.07, s.d. = 6.08) and the fricatives (mean = 1.09, s.d. = 0.81) as well as between the affricate (mean = 11.13, s.d. = 5.14) and the fricatives are significant, and as expected, differences between the plosives and the affricate are not significant.

TABLE I. Results for each consonant position.

Consonant position	Errors/total Fricative	Errors/total Plosive and affricate	Percent correct
Initial (training)	7/200	1/160	97.8
Initial (testing)	11/200	4/160	95.8
Medial	0/180	11/144	96.6
Final	0/160	9/160	97.2
Combined	18/740	25/624	96.8

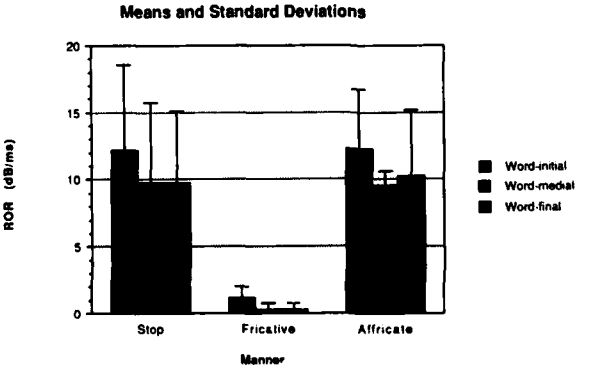


FIG. 7. A bar graph of ROR values (dB/ms) versus manner across the three word positions. These data indicate both that ROR values can be used for the manner distinction and that ROR values for word-medial and word-final positions are lower than those in word-initial position. Similarly, mean and standard deviations for each of the cases can be compared.

D. ROR versus sex

Differences in mean ROR values for males versus females were found to be significant only for the affricate (at $\alpha = 0.01$). For the affricate, males had a mean ROR value of 9.81 dB/ms, while the corresponding female value was 12.46 dB/ms. The only other significant difference was for plosives in word-initial position where average ROR values were 12.17 dB/ms for males and 11.22 dB/ms for females.

E. ROR versus vowel

Average ROR values across vowels exhibited no significant differences. Only in word-medial position was a significant difference observed where mean peak ROR values for all /i/'s (3.91 dB/ms) were lower than those for /a/ (4.8 dB/ms) and /u/ (4.49 dB/ms).

IV. DISCUSSION

The algorithm presented here, based on only three measures, provides an accurate method for distinguishing plosives and affricates from fricatives. While various combinations of other measures may yield similar results, an implementation based on log rms energy, ROR, and zero-crossing rate is straightforward and provides consistent results. Discussions of the development of the algorithm, of mislabelings, of results for speakers, and of ROR as a perceptual cue follow.

A. Algorithm design

Candidate measures were selected that could be easily and quickly obtained. For example, log rms energy measure which has been commonly used for both broad categorization of speech sounds and end-point detection (e.g., Rabiner and Sambur, 1975; Weinstein *et al.*, 1975; Lamel *et al.*, 1981) is one such measure. The derivative of log rms energy, defined here as rate of rise, is a natural candidate for a measure that may capture the often noticed differences between the initial rises of plosives and affricates and those of fricatives. The ROR information is supplemented by zero-crossing rates, allowing the algorithm to discard peaks related to nonspeech sounds or to vowel onsets. This is accomplished by evaluating a period of time around each ROR peak to find whether the number of zero crossings per second is remaining constant or increasing, thus correlated to consonant onset, and whether the number of zero crossings per second has reached a voiceless speech threshold during that period. The ROR information is also supplemented with information from the energy contour. The duration of an energy pulse is examined to ensure that it is large enough in both amplitude and duration to be considered a speech sound. Further, establishing a relative increase in level which the energy contour must exceed aids in rejection of vowel onsets following plosives or fricatives and biases the algorithm toward classifying low-valued peaks as not significant.

B. Fricative mislabeling

The mislabelings that occur using the algorithm discussed here are primarily related to nonspeech sounds (i.e., lip smacks and saliva pops) occurring at the onset (on the initial rise) of word-initial fricatives. The algorithm will identify these smack/fricatives as plosives; even visually, the energy contour, ROR, and zero-crossing rate are seemingly indistinguishable from a plosive, making it difficult to discard this kind of peak. This type of error accounts for at least 15 of the 18 word-initial fricative errors. Included in these 15 errors are those caused by two nonspeech sounds occurring in immediate succession. In this case zero crossings can rise above 2000 Hz, pulse duration can exceed 50 ms, and log rms energy can exceed 12 dB above the dB level at the time the peak in the derivative occurs. Thus the derivative peak resulting from two consecutive nonspeech sounds (appearing in the energy contour as a double-lobed pulse) can meet each of the criteria for a "significant" peak and, with the rapid ROR characteristic of this nonspeech sound, be classified as a plosive.

If these fricatives, ones immediately preceded by nonspeech sounds, are perceived as plosives, then the algorithm performs similarly to the human in labeling the tokens as plosives. This is supported in the case of 5 of the 18 word-initial fricative errors wherein the fricatives were transcribed by a phonetician as plosives.

C. Plosive mislabeling

Since the algorithm is tuned to detect plosives (because a high criterion was set for the relative increase in log rms energy necessary to label the peak as significant) and discard as many peaks as possible, few plosives are mislabeled. Weak bursts in /p/-initial utterances account for all of the 23 plosive errors. The energy contour of these weak /p/ bursts seems to be visually indistinguishable from a fricative energy contour, with a slowly rising slope and thus a small ROR, making automatic detection difficult. The /t/ and /k/ tokens are labeled as plosives in all cases. The affricate is only mislabeled two times and in those cases because of a low peak ROR value occurring in word-medial or final position.

D. Individual versus group results

One method examined for reducing the number of mislabeled plosives and fricatives was to adjust the ROR thresholds for individual speakers. It was found that such adjustments did not significantly increase the accuracy of classification. Another possibility, as yet unexplored, is that varying the criteria used for selection of a "significant" peak for an individual speaker may improve scores.

Percent correct for individual speakers ranged from 91.1%–99.6% (Table II). The lowest scores (91.1% and 93.3%) were seen in the word-initial case from two females who produced more nonspeech sounds than did the other talkers.

E. Rate of rise as a perceptual cue

We consider it quite likely that the rate of rise of log rms energy, or an appropriate psychophysical transform such as the rate of rise of loudness, can serve as a perceptual cue for the plosive/fricative distinction and, possibly, as a supplementary cue for the place of articulation of plosives.

Perceptual experiments indicate that plosives and affri-

TABLE II. Results for individual speakers.

Speaker	Sex	Initial	Medial	Final	Overall
1	F	100.0	100.0	98.8	99.6
2	M	100.0	98.8	100.0	99.6
3	M	97.8	···	···	97.8
4	M	97.8	···	···	97.8
5	F	96.7	···	···	96.7
6	M	97.8	93.8	96.3	96.0
7	F	93.3	93.8	95.0	94.0
8	F	91.1	···	···	91.1

cates can be distinguished from fricatives within the first 30 ms after the onset of the sound. For example, Blumstein and Stevens (1980) have shown that listeners can distinguish the place of articulation of plosive consonants when presented with the first 20 ms of the burst. Also, Jongman (1989) has shown that most fricatives can be identified after listening to their initial 30–50 ms. Since intraplosive and intrafricative identification can occur with these short segments, it is plausible that the plosive/fricative distinction could be made as well. Indeed, the results of Jongman (1989) are that his listeners rarely mislabeled fricatives as plosives when only the initial 20 ms of fricative sounds were presented.

One can also argue that ROR can serve as a supplementary cue for the place of articulation of plosives. In our data ROR is smallest for /p/'s, intermediate for /k/'s, and largest for /t/'s. If one sets the upper limit for /p/'s at 9.0 dB/ms, the lower limit for /t/'s at 13.5 dB/ms, and bounds /k/'s between these two limits, then /p/'s can be recognized with 80% accuracy, /k/'s with 71% accuracy, and /t/'s with 73% accuracy. By this approach one can categorize the place of articulation of voiceless plosives with about 75% accuracy by ROR alone. If one only contrasts /p/'s and /t/'s, an ROR threshold of 9 dB/ms results in 82% correct classifications without the use of any spectral information. Recently, Ohde and Stevens (1983) reported that burst amplitude could influence the perception of a synthetic /p/–/t/ continuum. For ambiguous tokens, the higher burst amplitudes resulted in more /t/ responses, while the lower burst amplitudes resulted in more /p/ responses. It is possible that the higher burst amplitudes had larger RORs, while the lower burst amplitudes had lower RORs. This could occur if all stimuli started from the same noise floor, had the same rise time, and differed only in their maximum amplitude. If measurements of their stimuli confirm this possibility, then it may be that these listeners were using ROR to disambiguate the spectrally ambiguous tokens near the middle of the synthetic continuum.

While it is shown here that voiceless affricates can be distinguished from voiceless fricatives by their peak RORs, numerous studies (for examples, see Dorman *et al.*, 1979; Repp *et al.*, 1978) have been interpreted as showing that several variables can influence this distinction. Among these variables are those mentioned by Dorman (Dorman *et al.*, 1980, p. 404), "...the presence and duration of the silent closure interval, the release burst, the rise-time of the fricative noise and the duration of the fricative noise." Other evidence speaks against the ROR of the burst as being an important cue. Under certain circumstances, listeners will label stimuli as voiceless affricates or voiceless fricatives in the absence of the plosive bursts usually associated with affricates (Howell and Rosen, 1983). Furthermore, van Heuven (1979) has reinterpreted Gerstman's (1957) data to mean that frication duration accounts for human identification performance better than does frication rise time. All of the results cited above indicate that peak ROR may not play an important role as a perceptual cue for the voiceless affricate/fricative distinction. In apparent contrast to the literature cited, our observation of unmodified natural speech, albeit in citation syllables recorded in excellent conditions, is that the peak RORs associated with the release burst[2] distinguish the affricate from the fricative, as can be seen in Fig. 7. In our opinion, peak ROR may be one of the more important cues for the voiceless affricate/fricative distinction in open set listening to natural speech. As past work has never measured peak ROR, we suggest that future work should include this measurement to aid in distinguishing among alternative hypotheses.

Finally, it is noted that, among the potential cues for the plosive/fricative distinction listed above and in the Introduction, ROR, as defined here, offers certain advantages. Unlike amplitude, ROR is invariant with amplification. Unlike rise time, it does not require the precise location of the onsets and terminations of periods of rise. And, unlike duration, ROR seems to be invariant with speech rate. It is our plan to design experiments to explicate the role of ROR in the perception of the manner distinction between plosives and fricatives and the place of articulation distinction for plosives. Preliminary analysis suggests that ROR may be a decisive factor only for certain reasonable combinations of spectra, amplitudes, and durations of the burst-friction components of plosives, affricates, and fricatives. Therefore, it is likely that ROR functions as part of a complex set of cues used by listeners in perceiving differences in manners of articulation between plosives and fricatives and in differences in place of articulation in plosives.

V. SUMMARY

Accurate categorical classification of voiceless plosives (including affricates) and fricatives can be achieved using only three measures: log rms energy, rate of rise of log rms energy, and zero-crossing rate. The ROR value is used as the primary classifier, with energy and zero-crossing rate being used primarily to discard spurious peaks, for example, those related to nonspeech sounds preceding the utterance or to vowel onsets. Furthermore, it is suggested that ROR, or its appropriate psychophysical transform such as the rate of rise of loudness, may serve as an important cue for the plosive/fricative distinction, and that it may serve as a contributing cue for place of articulation of plosives. Work currently in progress will test the merit of this approach for the classification of the voiced plosives, affricate, and fricatives.

ACKNOWLEDGMENTS

The authors wish to thank Joan Sereno and Allard Jongman for useful comments, information, and transcriptions,

Ira Hirsh for helpful discussions, and Robert Gilkey for suggesting improvements to the paper. We also wish to thank Stuart Rosen and an anonymous referee for their careful and thoughtful reviews of the manuscript. The research was supported by grants from the Air Force Office of Scientific Research (AFOSR-86-0335) and the National Institute on Deafness and other Communicative Disorders (R01-DC00296) to the Central Institute for the Deaf.

[1] Since it was heuristically determined that approximately 50 ms of information was sufficient for determining whether an ROR peak was significant, the 49-ms period following the peak was examined, thus providing 50 ms of information.

[2] Rapid rates of rise, with an average of 11.1 dB/ms (s.d. = 5.14), have been seen here for the affricate. Studies have shown that affricate rise times are shorter than fricative rise times (Gerstman, 1957; Cutting and Rosner, 1974; Howell and Rosen, 1983). Howell and Rosen considered, as is generally done, the rise time of the frication part of the affricate, ignoring the burst, and found rise times for affricates to have a mean of 33 ms (a mean of 76 ms for the fricatives). In the study presented here, however, the burst is included, thus explaining the rapid rates of rise found for the voiceless affricate.

Blumstein, S. E., and Stevens, K. N. (**1980**). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.

Carterette, E. C., and Jones, M. H. (**1974**). "On the statistics of spoken American English," *Proceedings Speech Communication Seminar*, Stockholm, Vol. 3, pp. 165–173.

Cutting, J. E., and Rosner, B. S. (**1974**). "Categories and boundaries in speech and music," Percept. Psychophys. **16**, 564–570.

Cutting, J. E., and Rosner, B. S. (**1976**). "Discrimination functions predicted from categories in speech and music," Percept. Psychophys. **20**, 87–88.

Dorman, M. F., Raphael, L. J., and Isenberg, D. (**1980**). "Acoustic cues for a fricative-affricate contrast in word-final position," J. Phon. **8**, 397–405.

Dorman, M. F., Raphael, L. J., and Liberman, A. M. (**1979**). "Some experiments on the sound of silence in phonetic perception," J. Acoust. Soc. Am. **65**, 1518–1532.

Gerstman, L. J. (**1957**). "Perceptual dimensions for the friction portions of certain speech sounds," Ph.D. thesis, New York University.

Halle, M., Hughes, G. W., and Radley, J. (**1957**). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. **29**, 107–116.

Heinz, J. M., and Stevens, K. N. (**1961**). "On the properties of voiceless fricative consonants," J. Acoust. Soc. Am. **33**, 589–596.

Howell, P., and Rosen, S. (**1983**). "Production and perception of rise time in the voiceless affricate/fricative distinction," J. Acoust. Soc. Am. **73**, 976–984.

Jongman, A. (**1989**). "Duration of frication noise required for identification of English fricatives," J. Acoust. Soc. Am. **85**, 1718–1725.

Lamel, L. F., Rabiner, L. R., Rosenberg, A. E., and Wilpon, J. G. (**1981**). "An improved endpoint detector for isolated word recognition," IEEE Trans. Acoust. Speech Signal Process. **ASSP-29**, (4), 777–785.

Ohde, R. N., and Stevens, K. N. (**1983**). "Effect of burst amplitude on stop consonant place of articulation," J. Acoust. Soc. Am. **74**, 706–714.

Oppenheim, A. V., and Schafer, R. W. (**1975**). *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).

Paliwal, K. K. (**1984**). "Effect of preemphasis on vowel recognition performance," Speech Commun. **3**(1), 101–106.

Paliwal, K. K., and Rao, P. V. S. (**1977**). "Acoustic phonetic recognition of continuous speech," in Proceedings of the 9th International Congress of Acoustics, Spain.

Paliwal, K. K., and Rao, P. V. S. (**1982**). "Synthesis-based recognition of continuous speech," J. Acoust. Soc. Am. **71**, 1016–1024.

Paliwal, K. K., Sinha, S. S., and Agarwal, A. (**1983**). "An isolated word recognition system for Hindi digits using linear time normalization," J. Inst. Electron. Telecommun. Eng. **29**(1), 18–22.

Pickett, J. M. (**1980**). *The Sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception* (University Park, Baltimore).

Rabiner, L. R., and Sambur, M. R. (**1975**). "An algorithm for determining the endpoints of isolated utterances," Bell Syst. Tech. J. **54**(2), 297–315.

Reddy, D. R. (**1966**). "Segmentation of speech sounds," J. Acoust. Soc. Am. **40**, 307–312.

Regel, P. (**1982**). "A module for acoustic-phonetic transcription of fluently spoken German speech," IEEE Trans. Acoust. Speech Signal Process. **ASSP-30**, 440–450.

Repp, B. H. (**1984**). "Closure duration and release burst amplitude cues to stop consonant manner and place of articulation," Lang. Speech **27**, 245–254.

Repp, B. H., Lieberman, A. M., Eccardt, T., and Pesetsky, D. (**1978**). "Perceptual integration of acoustic cues for stop, fricative, and affricate manner," J. Exp. Psychol: Human Percept. Perform. **4**, 621–636.

Schwartz, R., and Makhoul, J. (**1975**). "Where the phonemes are: Dealing with ambiguity in acoustic-phonetic recognition," IEEE Trans. Acoust. Speech Signal Process. **ASSP-23**(1), 50–53.

Shadle, C. H. (**1985**). "The acoustics of fricative consonants," Ph.D. thesis, MIT, Cambridge, MA.

Stevens, K. N. (**1980**). "Acoustic correlates of some phonetic categories," J. Acoust. Soc. Am. **68**, 836–842.

Strevens, P. (**1960**). "Spectra of fricative noise in human speech," Lang. Speech **3**, 32–49.

Upton, H. W. (**1968**). "Wearable eyeglass speechreading aid," Gallaudet Conference on Speech Aids, Am. Ann. Deaf **113**(2), 222–229.

van Heuven, V. J. (**1987**). "Reversal of the rise-time cue in the affricate-fricative contrast: An experiment on the silence of sound," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Nijhoff, Boston), pp. 181–187.

van Heuven, V. J. (**1979**). "The relative contribution of rise time, steady time, and overall duration of noise bursts to the affricate-fricative distinction in English: A re-analysis of old data," in *ASA-50 Speech Communication Preprint Experiment*, edited by J. J. Wolf and D. H. Klatt (Acoustical Society of America, New York), pp. 307–310.

Weinstein, C. J., McCandless, S. S., Mondshein, L. F., and Zue, V. W. (**1975**). "A system for acoustic-phonetic analysis of continuous speech," IEEE Trans. Acoust. Speech Signal Process. **ASSP-23**(1), 54–67.

1-18-91

Plosive/fricative distinction: The voiced case

LaDeana F. Weigelt

Steven J. Sadoff

James D. Miller

Central Institute for the Deaf

818 South Euclid Ave.

St. Louis, MO 63110

Original copy

Submitted to: Speech Communication

Received:

AFOSR Grant G-AFOSR-86-0335
Final Technical Report
Appendix

G

Abstract. We previously reported an algorithm which distinguishes voiceless plosives from voiceless fricatives with a success rate of 96.8% [J. Acoust. Soc. Am. In press]. A similar, but modified, algorithm also makes this distinction for the voiced case. Here results are presented on the modified algorithm. The input signal is high-pass filtered and two measures of the resulting waveform are used. One measure is the log rms energy. The other is the derivative of log rms energy over time or rate of rise (ROR) of amplitude. Peaks in the ROR contour are ~~considered~~ taken into account for significance to the plosive/fricative distinction by examining the log rms energy. Then, the magnitude of the first significant peak in the ROR contour is used as the primary classifier. The resulting algorithm was tested on 1128 tokens (with the consonant of interest in word-initial, medial or final position) recorded in an anechoic chamber with 564 tokens serving as a training set and the remaining data serving as a test set. The overall success rate was 96.3%. This new algorithm, developed for the voiced case, was also applied to the previously studied voiceless data. The rate of correct classification across all word positions within the word from both the voiced and voiceless cases was 95.8%.

Zusammenfassung. Abstract in German.

Résumé. Abstract in French.

INTRODUCTION

In a previous paper an algorithm was described which distinguished voiceless plosives from voiceless fricatives using the rate of rise (ROR) of the logarithm of the root-mean-square (rms) energy of the acoustic waveform and achieved a rate of 96.8% correct classification (Weigelt et al., In press). The purpose of this paper is to describe the continuation of that work which has resulted in development of a method, similar to that used in the voiceless case, for automatically distinguishing voiced plosives from voiced fricatives again based on ROR. Further, results are presented on the more general case of the plosive/fricative distinction where the consonant can be either voiced or voiceless.

Possible applications of an algorithm to distinguish plosives from fricatives are wearable eyeglass speechreading aids like that described by Upton (1968) and automatic speech recognition systems (Reddy, 1966; Weinstein et al., 1975; Schwartz and Makhoul, 1975; Paliwal and Rao, 1977; Regel, 1982; Paliwal and Rao, 1982) both of which classify speech segments into broad phonemic categories (e.g. stop-like or fricative-like). It is also a part of the general effort in our laboratory to develop a theory of phonetic perception (Miller, 1989; Miller, 1987).

While investigators have postulated some acoustic correlates of this distinction, no reliably detectable correlate has been systematically shown to accurately distinguish these classes of sounds. While the voicing distinction for plosives has been investigated extensively (e.g. Lisker and Abramson, 1964; Klatt, 1975; Kuhl and Miller, 1975) as has the voiceless affricate/fricative distinction (e.g. Gerstman, 1957; Howell and Rosen, 1983) , less attention has been given to the cues for the voiced plosive/fricative

distinction. A discussion of the general properties of voiced plosives (Halle et al., 1957; Strevens, 1960) and voiced fricatives (Strevens, 1960; Hughes and Halle, 1956; Stevens et al., 1987) can be found in the literature. In general, plosives are often distinguished from other consonants by the presence of silences or transient bursts (Halle et al., 1957; Borden and Harris, 1984; O'Shaughnessy, 1987), and, similarly, fricatives are often distinguished from other consonants by the presence of high frequency noise of "sufficient" duration (Borden and Harris, 1984; O'Shaughnessy, 1987). However, it should be noted that durational differences used in synthesis for voiced plosives and voiced fricatives are minimal (Klatt, 1979) .

A summary of differences between the voiced plosives and fricatives, from which can be derived possible cues for their distinction, is provided by Pickett (1980) . One of the spectral cues discussed is differences in the murmur spectrum: the murmur spectrum of a voiced plosive is composed of very low frequencies whereas that for a voiced fricative is strong in the high frequencies (greater than 2.5 kHz). Also included in spectral cues are differences in the formant transitions in adjacent vowels where short formant transitions are seen to correspond to voiced plosives and formant transitions which are "somewhat longer" than plosives correspond to the voiced fricatives. There are also differences in articulation: voiced plosives have a brief period of closure and voiced fricatives do not. Finally, voiced plosives tend to have abrupt spectral changes with a brief release transient whereas voiced fricatives have a spectrum which rather continuously changes from simply periodic to amplitude-modulated random turbulence with no release transient.

The method discussed in this paper for distinguishing plosives and fricatives takes advantage of the fact that voiced plosives (the voiced affricate is treated as a plosive

because it is plosive initial) have a brief burst whereas the voiced fricatives do not. The measure used to detect this difference is the ROR of log rms energy. The onset of a plosive or affricate is abrupt therefore having a high ROR value. The onset of a fricative, on the other hand, has a gradual rise in log rms energy and therefore a relatively low ROR value. Thus, an ROR threshold can be set such that phonemes with its first significant peak ROR value greater than the threshold are labeled as plosives and those with the ROR value of their first significant peak below the threshold are labeled as fricatives. While this abrupt amplitude change has been noted elsewhere (Stevens, 1980; Pickett, 1980), most earlier observations have been qualitative in nature.

In work reported here, for most cases the value of the highest peak in the ROR contour of a plosive or fricative is used for the plosive/fricative distinction. However, large ROR peaks resulting from non-speech sounds (i.e. lip smacks and saliva pops) preceding the syllable can be confused with consonant onset. For the application discussed here, these spurious peaks are rejected using methods similar to those commonly used for signal endpoint detection, i.e. setting thresholds for relative energy magnitude and energy pulse duration (Rabiner and Sambur, 1975; Lamel et al., 1981). It is also possible to confuse prevoicing with consonant onset. Word-initial voiced plosives are sometimes prevoiced by English speakers (Kuhl and Miller, 1975) as are those in intervocalic position (Lisker and Abramson, 1964). To remove ROR peaks corresponding to prevoicing, the acoustic waveform was highpass filtered with a cutoff frequency of 625 Hz.

Lisker & Abramson, 1967

I. METHODS

A. Speakers and Tokens

Eight native-Americans, four male and four female, with no known history of either speech or hearing disorders served as speakers. Of these speakers, only 2 had a formal course in phonetics, 2 had some exposure to phonetic transcription. and 3 had no exposure to phonetics. The word lists that were used were of the form VCV or CVC where the consonant of interest was a voiced consonant. Separate recordings were made for each word position. The word lists are ~~now~~ as follows described. When the consonant of interest is in word-initial position, 70 CVC tokens were used: 7 consonants (3 plosives and an affricate [b, d, g, d_3], 3 fricatives [z,v, ð]) X 10 vowels [i,I,ɛ,æ,a,ʌ,ɔ,U,u,ə] . When the consonant of interest was is word-medial position, 72 VCV tokens were used: 3 vowels [i,a,u] X 8 consonants [b,d,g,d_3,z,ž,v, ð .] X 3 vowels [i,a,u]. And, when the consonant of interest is in word-final position, 70 CVC tokens were used: 10 vowels X 7 consonants [b,d,g,~~z~~,v, ð].

B. Recordings

The recordings were made in an anechoic chamber using a low-noise microphone/ preamplifier combination (Bruel & Kjaer 4179/2660). The microphone was placed at a height equal to and 1/2 meter in front of the subject's mouth (zero degrees angle of incidence). In this set-up, turbulent noise generated by the breath flow over the microphone was not a factor. Conversational speech levels are used, i.e., 60 to 65 dBA to the microphone. The microphone output is channeled directly into a Sony PCM-

501ES digital audio recorder (16 bit mode) with a JVC 720 VCR serving as the storage medium.

Prior to the recording, speakers familiarized themselves with the list of randomized words. For the VCVs, talkers were instructed to place the emphasis on the second syllable. Speakers began reading tokens while an appropriate recording level was selected. When that recording level had been set, a calibration tone was recorded for that particular subject. The calibration tone consists of a 1 kHz sine wave generated by a Hewlett-Packard 3325A synthesizer and maintained at a constant output level of 69.5 mV at the input of the Sony digital audio recorder (equivalent to 70 dB SPL at the microphone).

C. Analysis

The recordings were played back from the Sony recorder in analog form and redigitized with proper anti-aliasing. They were digitized using a Micro Technology Unlimited Digisound-16 analog-to-digital and digital-to-analog converter and a custom hardware interface developed in-house. Digitization was performed at 20 kHz with 16-bit precision and edited to include the entire fricative or plosive portion of each token. The files were each notch filtered at 60 Hz to remove AC noise and stored on a MicroVAX II.

Productions from all eight speakers were used for the word-initial study. Two male and two female speakers from the word-initial study were used for the word-medial and word-final studies. In the word-initial case, the productions were divided into two sets: all utterances from two males and two females were chosen as the training set with the 280 tokens recorded from the remaining four speakers serving as a test set.

In the word-medial and final position cases, the data were divided into two sets as well: all productions from one male and one female were chosen as a training set with the ~~remaining~~ tokens of the two remaining subjects used only for testing. Note that no retraining was done after examining the test sets.

II. ALGORITHM

The algorithm was developed using a combination of measures that could be easily calculated and efficiently obtained. First, the input waveform is preemphasized using the first order difference equation, $y(i) = x(i) - 0.98x(i-1)$ where $x(i)$ is the ith input sample. This one-zero filter, $H(z) = 1 - 0.98z^{-1}$, acting as a differentiator, emphasizes the high frequency components where most of the acoustic cues for voiceless consonants reside. The use of preemphasis for consonant discrimination has been found to be of value (Paliwal, 1984).

The waveform is also highpass filtered at a cutoff frequency (cf) of 625 Hz to remove prevoicing murmurs. A cf of 625 Hz was chosen after testing highpass filters with a cutoff range of 125 to 9000 Hz. Across the training sets from each word position, the highest percentage correct (using the best ROR threshold for each data set) was attained when using a cf of 500 or 625 Hz (Fig. 1). While performance at neighboring cfs was similar, the value of 625 Hz was chosen for two reasons. ~~First, the optimal threshold range for that cf included~~ 2.24 dB/ms, ~~the~~ ROR threshold used for optimal performance on the voiceless plosive/fricative distinction. Second, when using one ROR threshold for both word-medial and final positions, the highest percent correct was obtained with a cf of

625 Hz.

The short term log rms energy, E_n, is defined as

$$E_n = 10 \log \left((1/N) \sum_{i=n}^{n+N-1} [y(i)w(i-n)]^2 \right) \quad (1)$$

where $w(i)$ is the standard finite duration Hamming window (Oppenheim and Schafer, 1975, page 242):

$$w(i) = 0.54 - 0.46 \cos \left(\frac{2\pi i}{N-1} \right) \quad 0 \leq i < N. \quad (2)$$

For the Hamming window a value of $N = 480$ samples (24 ms frames) is used. This window is moved in 20 sample (1 ms) steps. Notice that throughout this paper, indices within parentheses refer to sample number while subscripts refer to frame number. The primary measure used is the Rate Of Rise of log rms energy, ROR, which is defined as

$$ROR = (E_n - E_{n-1})/(\Delta t)$$

where Δt is the separation (in seconds) of the energy measurements. In our work, $\Delta t = (20 samples)/(20000 samples/sec) = 0.001$ seconds.

Peaks are located in the ROR function and are sorted in descending order by amplitude. Only peaks with positive ROR values need to be considered since peaks with negative ROR values correspond to offset rather than onset of the consonant. Each of these peaks are considered in order (largest peak first) as a potential candidate. If a candidate peak becomes significant if it passes each of the criteria to be described below and justified in Section IV-A, then a significant peak is found. If a peak fails any of the criteria, then it is discarded as being non-significant. The rate of rise value of the first significant peak is used as the peak ROR value for that token. Sounds with peak ROR values above a

predetermined cutoff are classified as plosives, while sounds with values below the cutoff are labeled as fricatives. If no significant peaks are found, the segment is labeled as a fricative.

In most cases the largest peak in the ROR contour corresponds to the onset of the plosive or fricative and thus is the "significant" peak. However, in some cases ROR peaks corresponding to non-speech sounds or vowel onsets can also have large values and thus can be confused with consonant onset. By examining training set data and visually choosing criteria and then testing these criteria, we were able to eliminate these spurious peaks. For a candidate ROR peak to be labeled as significant, then, the following conditions must be met. First, for the 49 ms period following the peak, the value of E_n must never fall below the value of E_n at the peak. Second, the maximum value of E_n for the following 49 ms must be at least 12 dB above the value of E_n at the peak. Since it was heuristically determined that approximately 50 ms of information was sufficient for determining whether an ROR peak was significant, the 49 ms period following the peak was examined thus providing 50 ms of information. While the exact values specified in the criteria may not be crucial and perhaps other combinations of measures may be equally effective, the criteria described below have proven to give excellent results.

III. RESULTS

Examination of algorithm performance on both the training and test sets supports the usefulness of using log rms energy and ROR for the plosive/fricative distinction.

The operation of the algorithm can be better appreciated by looking at easy-to-label examples of both a typical plosive and fricative as well as more difficult to label examples. A typical example of a plosive and a fricative waveform (where the waveform has been highpass filtered at 625 Hz) with log RMS energy, ROR and zero-crossings is shown in Figs. 2 and 3. In Fig. 2, a plosive, typical of those seen in these voiced data sets, can be seen to have a characteristic abrupt increase in log rms energy at the plosive onset and a corresponding sharp peak in the ROR contour. The "significant" peak, indicated by the arrow, has an ROR value well above the threshold and is correctly labeled, then, as a plosive. In Fig. 3, the typical voiced fricative is seen to have a gradual rise in log rms energy, rather than the abrupt rise seen for the plosive, and low peak values in the ROR contour. The "significant" peak, identified by the arrow, has an ROR value below threshold and, therefore, was correctly labeled as a fricative.

Two examples having log rms energy contours which may be difficult to label by visual inspection can be seen in Figs. 4 and 5. The plosive [b], seen in Fig. 4, has both a gradually rising energy contour and a relatively small increase in energy level at onset similar to that seen for a fricative. Also, it is difficult to label the token as a plosive or fricative based purely on its waveform. However, its ROR value is above the ROR threshold and thus is labeled correctly as a plosive. The fricative [z] seen in Fig. 5 has an energy contour similar to that seen for many plosives. However, its ROR value for the significant peak falls below the threshold value and is thus correctly labeled as a fricative. Thus cases which could be difficult to label can be correctly labeled by this algorithm.

A. ROR threshold

The ROR threshold for the plosive/fricative distinction was determined using the training sets from each position. The percentile plot of peak ROR values for the plosives, fricative and affricate (Fig. 6) provides a visual method of obtaining an ROR cutoff for the plosive/fricative distinction. Several possible ROR threshold values could be chosen thereby trading misclassifications of plosives with those of fricatives. The highest percentage of correct classifications for the training sets (across all word positions) is obtained when the ROR threshold value is between 2.166 and 2.237 dB/ms. Since it has been found that the plosives and affricate [b,d,g,dʒ] appear almost equally as often in conversational speech as do the fricatives [v,z,ž, ɣ :] (Carterette and Jones, 1974), the ROR cutoff was chosen at the midway point of this ambiguity region, 2.2 dB/ms.

In Fig. 7, peak ROR values separated by phoneme can be seen. These data indicate that not only can ROR be used to distinguish between plosives and fricatives but it may, in the case of plosives, also be used as a contributing cue to identify place of articulation. Highest ROR values for the plosives are obtained for [d] bursts, intermediate values for [g] bursts, and lowest ROR values are obtained for [b] bursts. This parallels the voiceless case where, among the plosives, highest ROR values were obtained for [t], intermediate for [k] and lowest for [p]. This correlates to mechanical pressure measures: of [p] and [t], the average peak pressure is greatest for [t] as is the pressure impulse (the integral of pressure over time) (Malecot, 1966). Of [d] and [b], the average peak pressure is greatest for [d] as is the pressure impulse (Malecot, 1966). Among fricatives, RORs are similar, however slightly higher for [z]'s and [ʃ]'s than the others.

B. Results for each consonant position

Across all word positions (including testing and training sets), the algorithm has a 96.3% rate of success (Table I) when using the threshold established using all of the training sets across all positions (2.2 dB/ms). The success rate on the word initial sets (testing and training) reached 96.4%. The plosive errors (2/240) were both low amplitude [b]s and the fricative errors (18/240) were due to relatively abrupt fricative onsets. There were no affricate errors. If the ROR threshold is adjusted to 2.63 dB/ms as indicated by performance on the word-initial training set alone, performance across the word-initial testing and training sets increases to 97.1%.

Performance on the word-medial data sets was the most accurate of the three word positions at 98.3%. The errors that did occur were three plosives (out of 108) and two on the affricate (out of 36) which had low amplitudes and low ROR values (one [b], two [g]s and two [d_3]s). If the ROR threshold is lowered to 1.417 dB/ms, the optimal threshold for the word-medial training set, the percent correct classifications rise to 99.7% (287/288) across the word-medial test and training sets.

In the word-final data sets, the percent success was 93.9% with 14 plosive errors (out of 120) and 3 affricate errors (out of 40). All of the plosive errors were on [b] bursts labeled as fricatives with four of those errors caused by bursts whose durations were less than 50 ms and the remaining [b] errors having low amplitude bursts or no burst at all. Each of the affricate errors was due to a low amplitude burst. If the ROR threshold is lowered to 1.94 dB/ms, the optimal threshold for the word-final training set, the percent correct classifications rise to 95.7% across word-final training and testing sets.

If the optimal threshold for each word-position is then used performance across data sets would be 97.4%. A system could be implemented which determines whether a particular plosive or fricative is in word-initial, medial or final position and then applies the appropriate threshold criterion. However, since ROR peak values in word-medial and final positions are similar, another possible solution would be to use two threshold criteria, one for word-initial position consonants and one for word-medial and final position consonants. If a priori knowledge that a consonant is or is not in initial position is assumed, and an ROR threshold for initial position of 2.48 dB/ms and 1.417 dB/ms for not initial position is used, then the overall success rate is 97.3%.

There are two possible explanations for the increase in percent correct brought about by lowering the ROR threshold for word-medial and final positions. One explanation is the fact that adjusting the ROR threshold specifically for individual data sets takes advantage of chance affects and therefore increases percentages correct by better predicting the error term. However, a more likely explanation is that plosives in word-medial and final position have, in general, a lower average ROR value than those in word initial position (especially when the word-initial token is spoken as an isolated syllable where more time is available for a build up of pressure behind the point of closure). Support for the latter hypothesis is provided in Fig. 8 where it can be seen that average ROR value for initial position voiced and voiceless plosives, affricates, and fricatives is much greater than that for medial and final positions. This implies that in an optimal implementation of this algorithm it may be useful to lower the ROR threshold when the consonant is in word-medial or final position.

C. ROR vs. manner

The results across all word positions were examined using a 3-factor ANOVA (ROR vs. phoneme, sex, and vowel). The differences in mean ROR values for the different manners of articulation are shown in Fig. 8 and are significant at $\alpha = .01$. In particular, using Tukey's procedure it was found that the differences between the means (both across all word positions and within each of the word positions) of the plosives and the fricatives as well as between the affricate and the fricatives are significant and, as expected, differences between the plosives and the affricate are not significant.

Also interesting are the significant differences between phonemes. From the data taken across all word positions it was found that the mean peak ROR value for [d] (12.11 dB/ms) is significantly greater than that for [d_3] (9.65 dB/ms), [b] (5.81 dB/ms) and, of course, the fricatives at $\alpha = .01$. Also at $\alpha = .01$, the mean peak ROR value for [g] (10.57 dB/ms) is significantly greater than that for [b] and for each of the fricatives. The mean peak ROR value for [b] is also significantly ($\alpha = .01$) greater than that for the fricatives. When an α level of .05 is accepted, the mean peak ROR value for [d] is significantly greater than that for [g].

D. ROR vs. sex

Differences in mean ROR values for males versus females were found to be significant across all word positions (female=6.27 dB/ms, male=5.07 dB/ms) as well as within word-initial (female=8.07 dB/ms, male=6.65 dB/ms) and word-medial position (female=5.59, male=3.75). For word-final position only the affricate mean ROR value for females (6.79 dB/ms) was found to be significantly greater than that for males (4.84

dB/ms). When examining data collected across all word-positions the mean peak ROR values for the females were significantly greater than that for males for each manner of articulation at $\alpha = .01$.

E. ROR vs. vowel

In looking at average ROR values across vowels, interactions were observed for word-initial and medial positions. When the consonant of interest was in word-initial or word-medial position there were significant interactions between the consonant and the vowel (the initial vowel in the word-medial case). In both word positions, significant differences between manners and places of articulation across each vowel were found. Also in word-medial position, for vowel and sex the no-interaction hypothesis was rejected. In this case across male speakers, mean peak ROR values for the consonants following [a] (4.84 dB/ms) were found to be significantly greater than those for [i] (3.04 dB/ms). Because of the already high percent performance of this algorithm, it is unlikely that by optimizing decision criteria for each of these significant interactions could significantly improve results.

F. Individual results

Percent correct for individual speakers ranged from 89% to 100% (Table II). The lowest percent correct can be seen for a speaker when the consonant of interest was in word-final position; each of those errors were low amplitude [b]s. If the ROR threshold is adjusted to 1.417 dB/ms, as suggested by the differences in optimal thresholds between word-initial and final positions, the performance for this speaker on this set would rise

to 95.7%.

IV. DISCUSSION

With the algorithm described here, based on ROR, voiced plosives can be accurately distinguished from voiced fricatives. While various combinations of other measures may yield similar results, this method, using only the log rms energy and derivative of log rms energy, is straight forward and provides consistent results. Discussions of motivation for and development of the algorithm, of mislabelings and individual results, of the results of applying the voiced algorithm to the voiceless case, and, finally, of ROR as a perceptual cue will follow.

A. Algorithm design

The algorithm was developed from the algorithm used for the voiceless case (Weigelt et al., ~~in press~~ 1990 JASA in press). Measures were selected which could be easily and quickly obtained and which had proven successful in other applications. For example, the log rms energy measure, which has been commonly used for both broad categorization of speech sounds and endpoint detection (e.g. Rabiner and Sambur, 1975; Weinstein et al., 1975; Lamel et al., 1981) , is one such measure. The derivative of log rms energy, defined here as Rate of Rise, is a natural candidate for a measure that may capture the ~~often-noticed~~ differences between the initial rises of plosives and affricates and those of fricatives. The ROR information is supplemented with information from the energy contour allowing the algorithm to discard peaks related to non-speech sounds. This is accomplished by examining energy lobes to insure that they are both large enough in amplitude and

duration to be considered a speech sound. Further, establishing a relative increase in level which the energy contour must exceed aids in rejection of vowel onsets following plosives or fricatives and biases the algorithm towards classifying low-valued peaks as not significant.

B. Mislabelings

The majority of the mislabelings that occur using the algorithm discussed here are simply a result of a fricative having a relatively high ROR or a plosive having a relatively low ROR. This is different than in the voiceless case where a majority of the mislabelings resulted from non-speech sounds occurring at the onset (on the initial rise) of word-initial fricatives. When the consonant of interest was in word-final position, 4 of the 17 plosive/affricate errors were the result of the peak in ROR associated with the consonant onset reaching less than 12 dB in relative energy level thus being classified as not significant. Four additional plosives had energy levels which did not remain above the energy level at the peak for 50 ms, thus being classified as not significant. These types of errors are difficult to eliminate. Both of these criterion are necessary to aid in the rejection of ROR peaks resulting from non-speech sounds with the first criterion being used to reject vowel onsets and bias the algorithm towards classifying low-valued peaks as not significant. The second criterion is necessary to eliminate non-speech sounds which are seen to be as long as 50 ms in duration. Since these errors made up a small percentage of the overall errors no new criteria were designed to eliminate those specific mislabelings.

All of the fricative errors occur in word-initial position. These mislabeled fricatives

have high ROR values like those associated with plosives. In the earlier study of the voiceless plosive/fricative distinction, it was found that some of the mislabeled fricatives were labeled by listeners as plosives (Weigelt et al., In press). Thus it was hypothesized that these voiced fricatives with high RORs may also be perceived as voiced plosives. To test this hypothesis, two experienced listeners were asked to listen to the 560 CVC tokens where the consonant of interest was in word-initial position and label the word-initial consonant as a plosive or fricative. However, only one fricative was mislabeled by a listener as being a plosive; this fricative was also mislabeled by the algorithm. Two plosives were mislabeled by a listener as being fricatives; one of these plosives was also mislabeled by the algorithm.

C. Relationship to voiceless plosive/fricative distinction

The algorithm applied to the voiced plosive/fricative data sets is similar to that used for the voiceless plosive/fricative case. The differences are in the pre-filtering of the waveform and in the use of zero-crossing criteria to determine whether a given ROR peak is significant. In the voiced case, prevoicing can be seen before the onset of plosives of fricatives which can be confused with consonant onset. Thus, for the voiced case, it was necessary to filter out the prevoicing. Also, whereas zero-crossing information was used in the voiceless case, such information was not used in the voiced case. In the voiced case, the zero-crossing criteria set up for the voiceless case, especially that requiring that the zero-crossings reach a voiceless threshold within 50 ms of the ROR peak, simply could not apply.

While it was clear that the algorithm developed for the voiceless case did not imme-

diately apply to the voiced case, it was hypothesized that the algorithm developed for the voiced case may succeed in making the voiceless plosive/fricative distinction. Thus, the voiced plosive/fricative distinction algorithm was applied to the voiceless data sets. Results can be seen in Table III. Using an ROR threshold of 2.24 dB/ms across all word positions from both the voiced and voiceless data sets, the percentage correct reaches 95.8%. If, however, as suggested earlier in this paper, the ROR threshold is lowered to 1.68 dB/ms (which was used also in the voiceless case for optimal scores for word-medial and final positions) for consonants in word-medial and final positions and an optimal threshold of 2.63 dB/ms for the word-initial case is used the overall percent correct rises to 96.7%.

The results obtained using one ROR threshold across the voiced and voiceless cases indicates that the voiced and voiceless plosives as well as the voiced and voiceless fricatives may have similar RORs; differences are small relative to the variance but they are consistent. Further support for this hypothesis is provided in Fig. 8 where it can be seen that only for final-position plosives and the affricate is there a significant difference between mean ROR values for the voiced and voiceless case. Although a voiceless plosive or fricative has a greater ROR on the average than the voiced, it is not a significant difference. This is interesting ~~in light of the fact that a differences between voiced and voiceless plosives include the fact that~~ because in the production of voiceless plosives there is more pressure built up behind the point of closure ~~than~~ for the voiced plosives (Halle et al., 1957; Malecot, 1966) and, similarly, that voiceless plosives are produced with more articulatory effort (a greater "force of articulation") than are voiced plosives (Slis, 1971; Malecot, 1970). These data suggest, then, that ROR is not purely dependent on

the amount of pressure built up before release.

In summary, we have developed two algorithms for distinguishing plosives (and affricates) from fricatives. One algorithm applies strictly to the voiceless case and the other can be applied to the voiced or voiceless case with both performing well for their specific task. The algorithm described in this paper is slightly less computationally intensive and can perform well for both the voiced and voiceless case and thus may be preferred.

D. Rate of rise (ROR) as a perceptual cue

Results obtained in this study correlate with those found in our voiceless study (Weigelt et al., In press) in ~~indicating~~ suggesting that the Rate of Rise of log rms energy, or an appropriate psychophysical transform such as the rate of rise of loudness, ~~can serve~~ might be used as a perceptual cue for the plosive/fricative distinction and, possibly, as a supplementary cue for the place of articulation of plosives. In support of the former hypothesis is the fact that ROR can be used to distinguish both voiced and voiceless plosives from voiced and voiceless fricatives. In support of the latter hypothesis, it was found that if [b] can be distinguished from [d] and [g] with 70% accuracy. Contrasting only [b]s and [d]s, an ROR threshold can be set at 8.77 dB/ms and this distinction can be made with 77% accuracy.

While it is shown here that affricates (as well as plosives) can be distinguished from fricatives by their peak RORs, other studies (Dorman et al., 1979; Repp et al., 1978; Howell and Rosen, 1983; van Heuven, 1979) have been interpreted as showing that several variables can influence this distinction. Our observations of unmodified natural speech, albeit in

citation syllables recorded in excellent conditions, is that the peak RORs associated with the release burst well distinguish the affricate from the fricative.

It is noted, then, as it was for the voiceless plosive/fricative distinction (Weigelt et al., In press), that among the potential cues for the plosive/fricative distinction ROR, as defined here, offers certain advantages. ROR is invariant with amplification, does not require the precise location of the onsets and termination of periods of rise, and seems to be invariant with speech rate. Therefore, in our opinion, it is likely that ROR functions as a part of a complex set of cues used by listeners to perceive differences in manners of articulation between plosives and fricatives and differences in place of articulation for plosives. It is recommended that this measure should be included in future work to aid in distinguishing among alternative hypotheses.

V. SUMMARY

Accurate categorical classification of plosives and fricatives can be achieved using only two measures: log rms energy and Rate of Rise of log rms energy. The ROR value is used as the primary classifier, with relative energy increases and energy pulse durations being used to discard spurious peaks, for example those related to non-speech sounds preceding the utterance. Furthermore, it is suggested that ROR, or its appropriate psychophysical transform such as the rate of rise of loudness, may serve as an important cue for the plosive/fricative distinction, and that it may serve as a supplementary cue for place of articulation of plosives.

VI. ACKNOWLEDGEMENTS

This research was supported by Grants from the Air Force Office of Scientific Research (AFOSR-86-0335) and the National Institute on Deafness and other Communicative Disorders (RO1-DC00296) to the Central Institute for the Deaf.

Table I: Results for each consonant position.

CONSONANT POSITION	ERRORS/TOTAL		% CORRECT
	FRICATIVE	PLOSIVE & AFFRICATE	
INITIAL	18/240	2/320	96.4
MEDIAL	0/144	5/144	98.3
FINAL	0/120	17/160	93.9
COMBINED	18/504	24/624	96.3

Table II: Results for individual speakers.

Speaker	Sex	Initial	Medial	Final	Overall
1	M	98.5	–	–	98.5
2	M	97.1	97.2	100.0	98.1
3	F	97.1	100.0	95.7	97.6
4	M	97.1	–	–	97.1
5	F	95.7	–	–	95.7
6	F	97.1	98.6	91.4	95.7
7	M	95.7	97.2	88.6	93.9
8	F	92.3	–	–	92.3

Table III: Combined results for voiced and voiceless data sets.

CONSONANT POSITION	ERRORS/TOTAL		% CORRECT
	FRICATIVE	PLOSIVE & AFFRICATE	
INITIAL	56/640	6/640	95.2
MEDIAL	3/324	13/288	97.4
FINAL	0/280	27/320	95.5
COMBINED	59/1244	46/1248	95.8

Figure 1. Percent correct across training sets from each word-position for different cutoff frequencies (cf) of the highpass filter used to remove prevoicing. The optimal ROR threshold value was used at each cutoff frequency to obtain the percent correct scores. A cf of 625 Hz was chosen for two reasons: (1) the optimal threshold range across all positions for that cf included 2.24 dB/ms, the ROR threshold used for the voiceless case, and (2) When using one ROR threshold for both word-medial and final positions, the highest percent correct was obtained with a cf of 625 Hz.

Figure 2. An example of a plosive, the [g] from the token [gUt], typical of the plosives seen in this study and correctly identified by the algorithm as a plosive. The waveform (highpass filtered at 625 Hz), log rms energy, rate of rise (ROR) of log rms energy and zero crossings are shown. Notice the rapid ROR of log rms energy at the plosive onset signaled by the peak in the ROR plot. The peak denoted by the arrow is the "significant" peak, the peak used to label the token as plosive or fricative. The apparent discrepancy between the position of the plosive onset in the waveform and the peak in the ROR contour results from the fact that the window used for the calculation of the log rms energy is defined in terms of its leftmost point.

Figure 3. An example of a typical fricative, the [v] from the token [vɛt], its waveform, log rms energy and ROR, correctly identified by the algorithm as a fricative. Notice the gradual rise of the energy contour correlated with a small peak (see arrow) in the ROR plot.

Figure 4. An example of a plosive, the initial plosive in [buθ], whose log rms energy looks more similar to a fricative than that of a plosive. With the ROR threshold set at 2.24 dB/ms, however, the ROR value of the significant peak allows it to be correctly

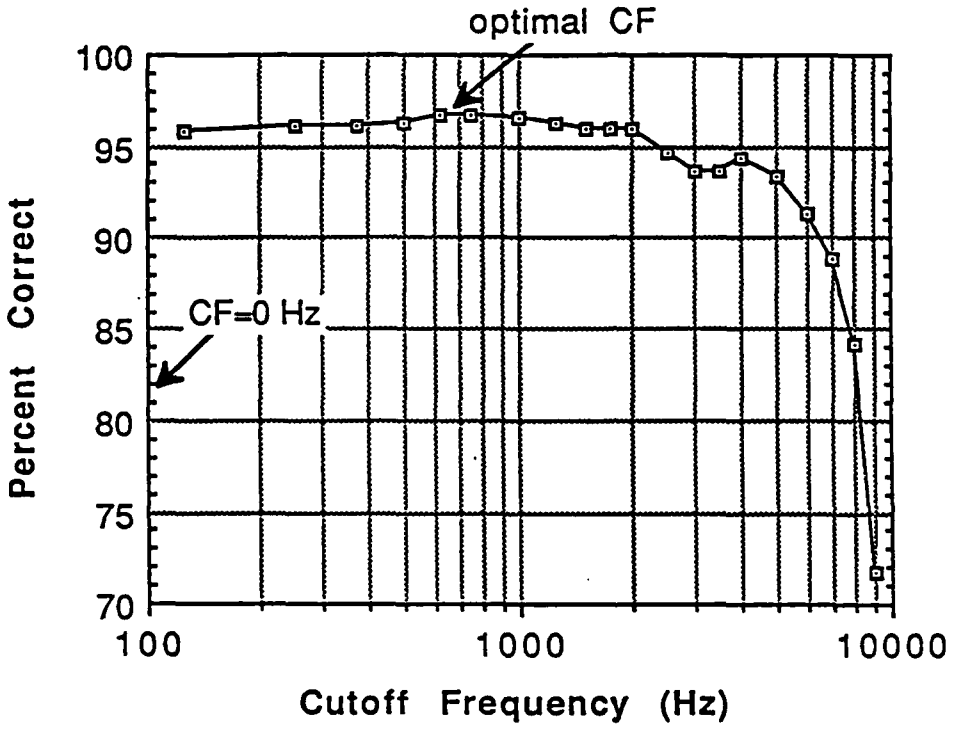
labeled as a plosive.

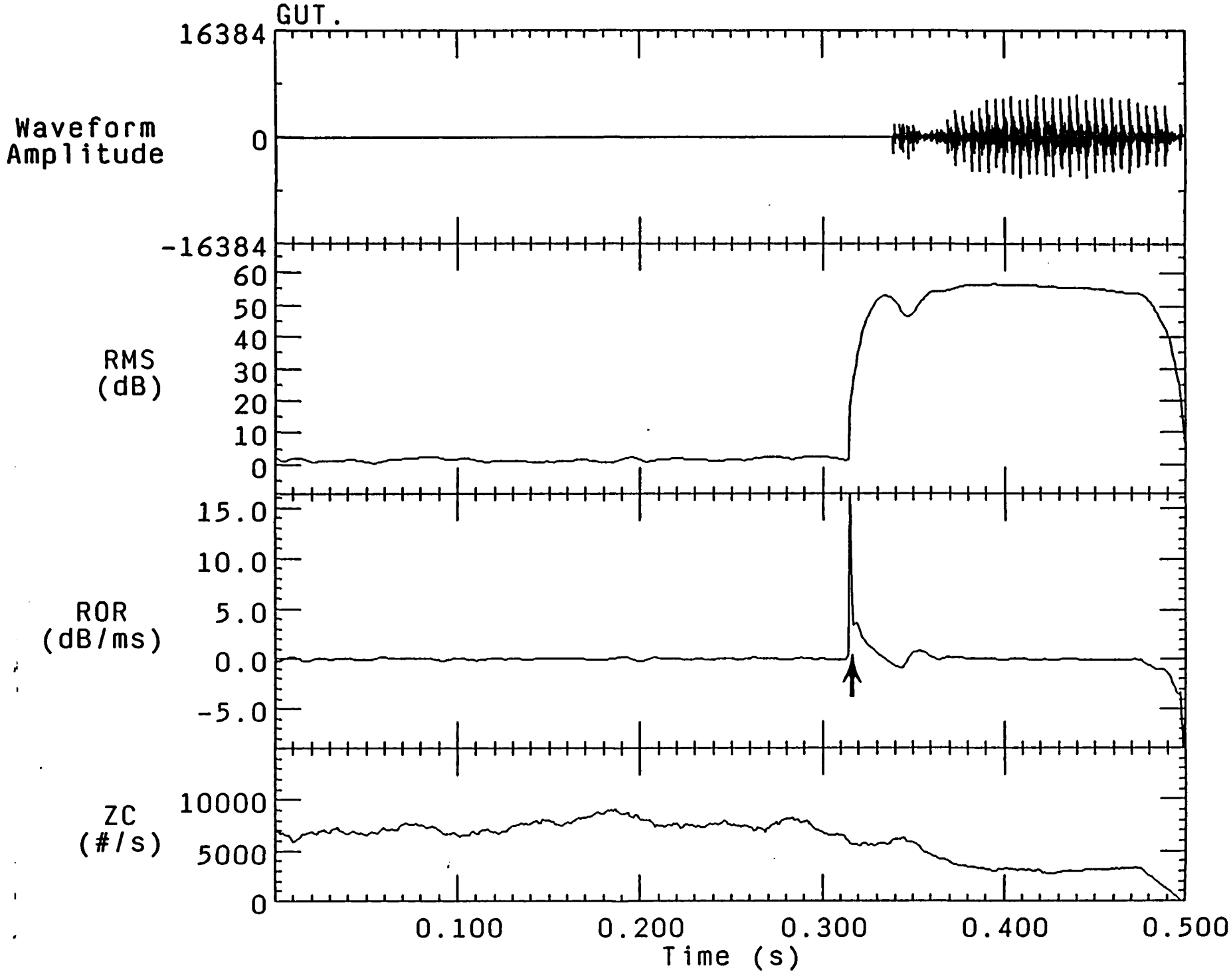
Figure 5. An example of a fricative, [z] from [zut], whose energy contour is similar to that of a plosive. With the ROR threshold set at 2.24 dB/ms, however, it is correctly labeled as a fricative.

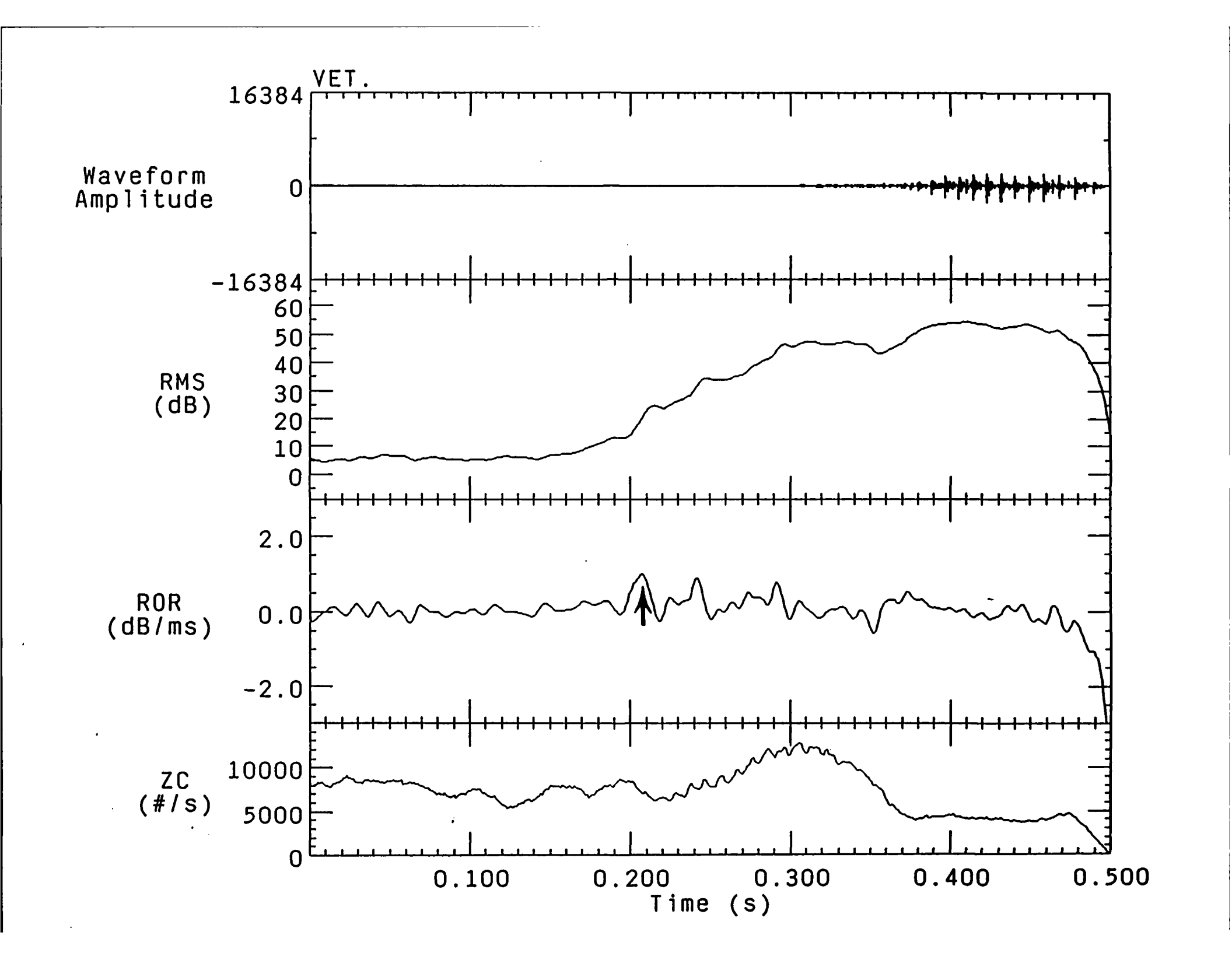
Figure 6. A plot of ROR values in dB/ms versus percentile ranks separated by manner of articulation for the 560 word-initial tokens. It can be seen that plosives and affricates have higher ROR values than do fricatives. Percent correct scores resulting from each choice of an ROR threshold can also be obtained.

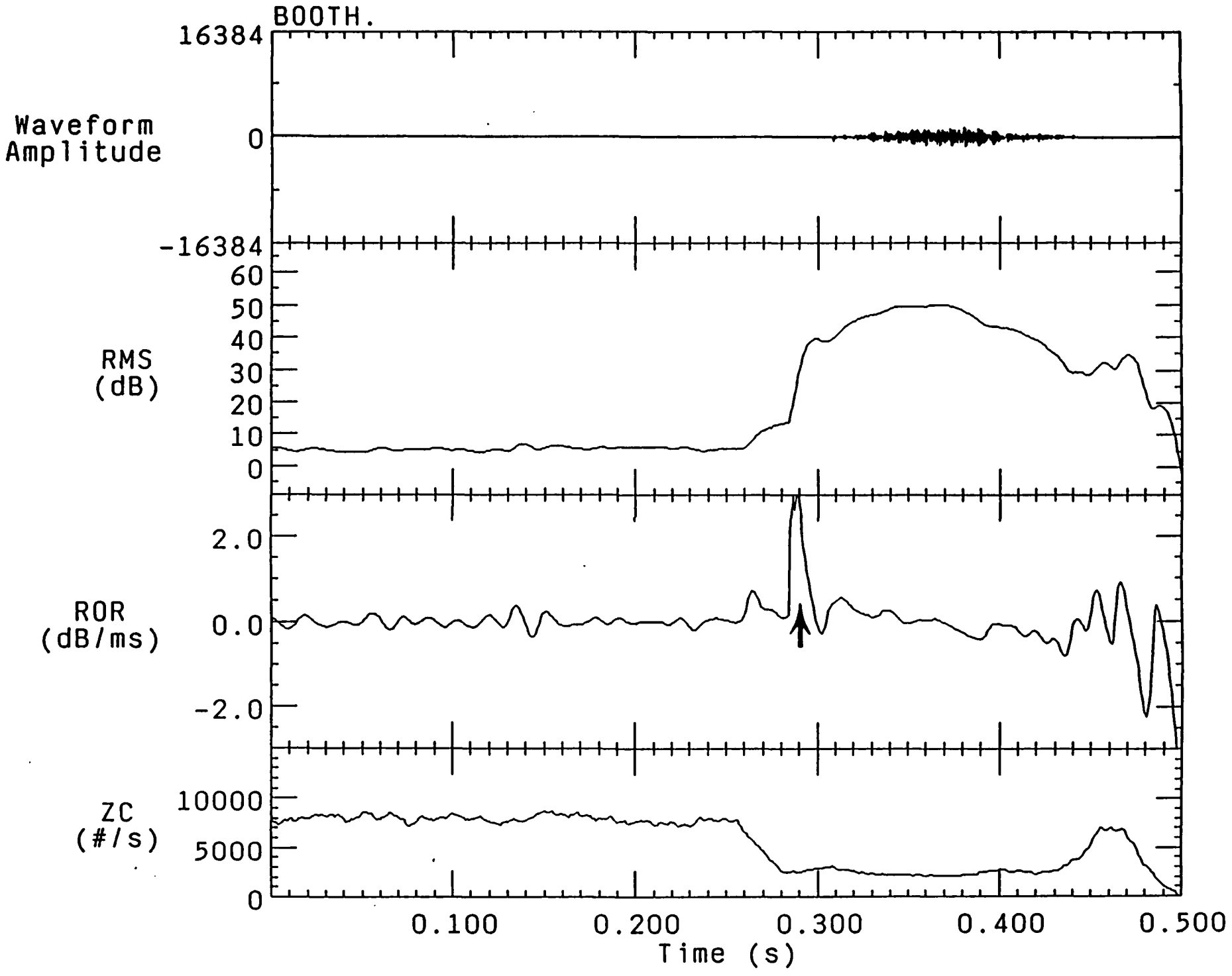
Figure 7. A plot of ROR values (dB/ms) versus percentile ranks for 560 word-initial tokens separated by place of articulation. It can be seen that of the voiced plosives, [d] has the highest ROR values and [b] the lowest. While the fricative ROR values tend to cluster, [z] has the highest ROR values of the group.

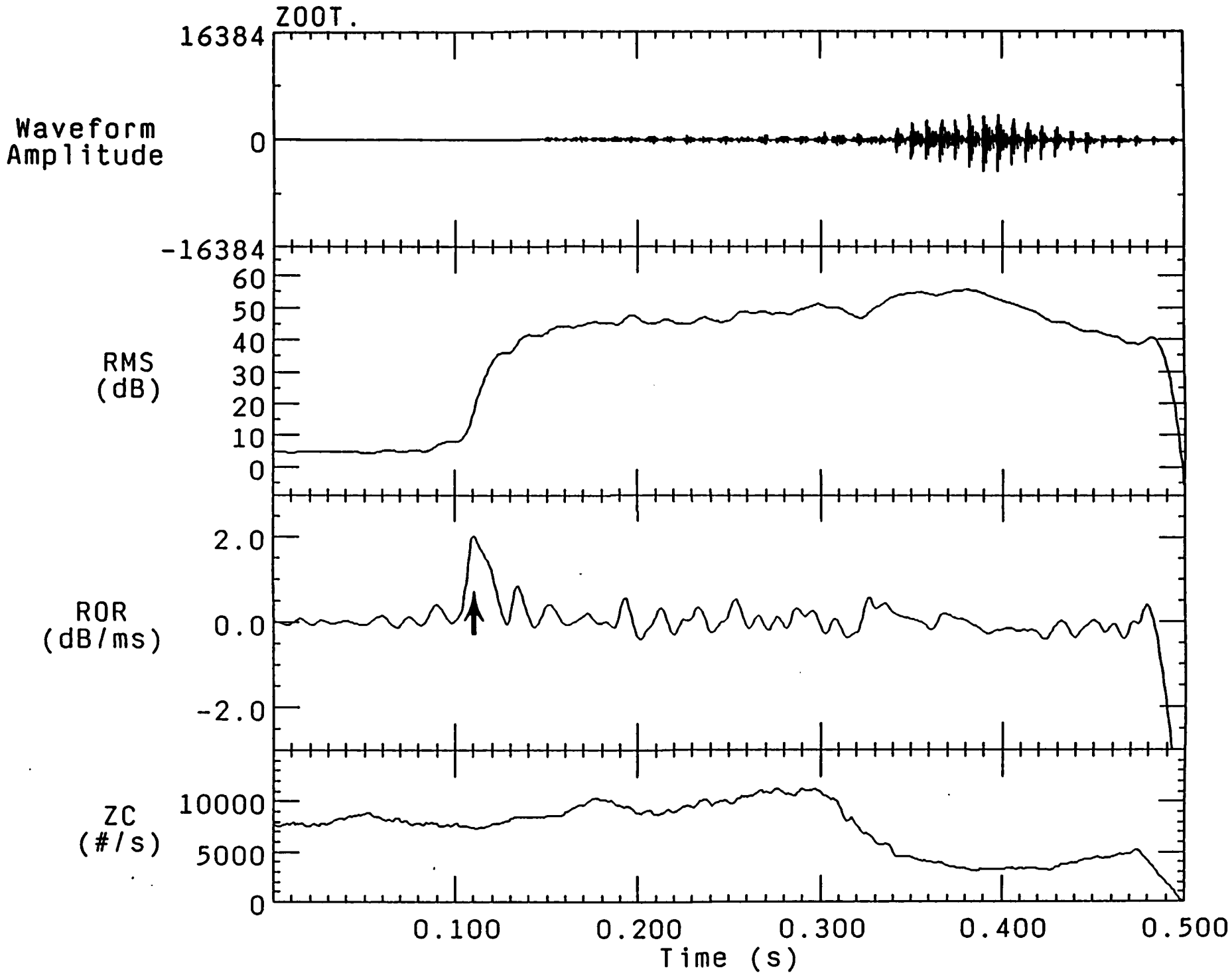
Figure 8. A bar graph of mean ROR values (dB/ms) versus manner across the three word positions for the voiced and voiceless cases. These data indicate both that ROR values can be used for the manner distinction and that ROR values for word-medial and word-final positions are lower than those in word-initial position. Similarly, mean and standard deviations for each of the cases can be compared.

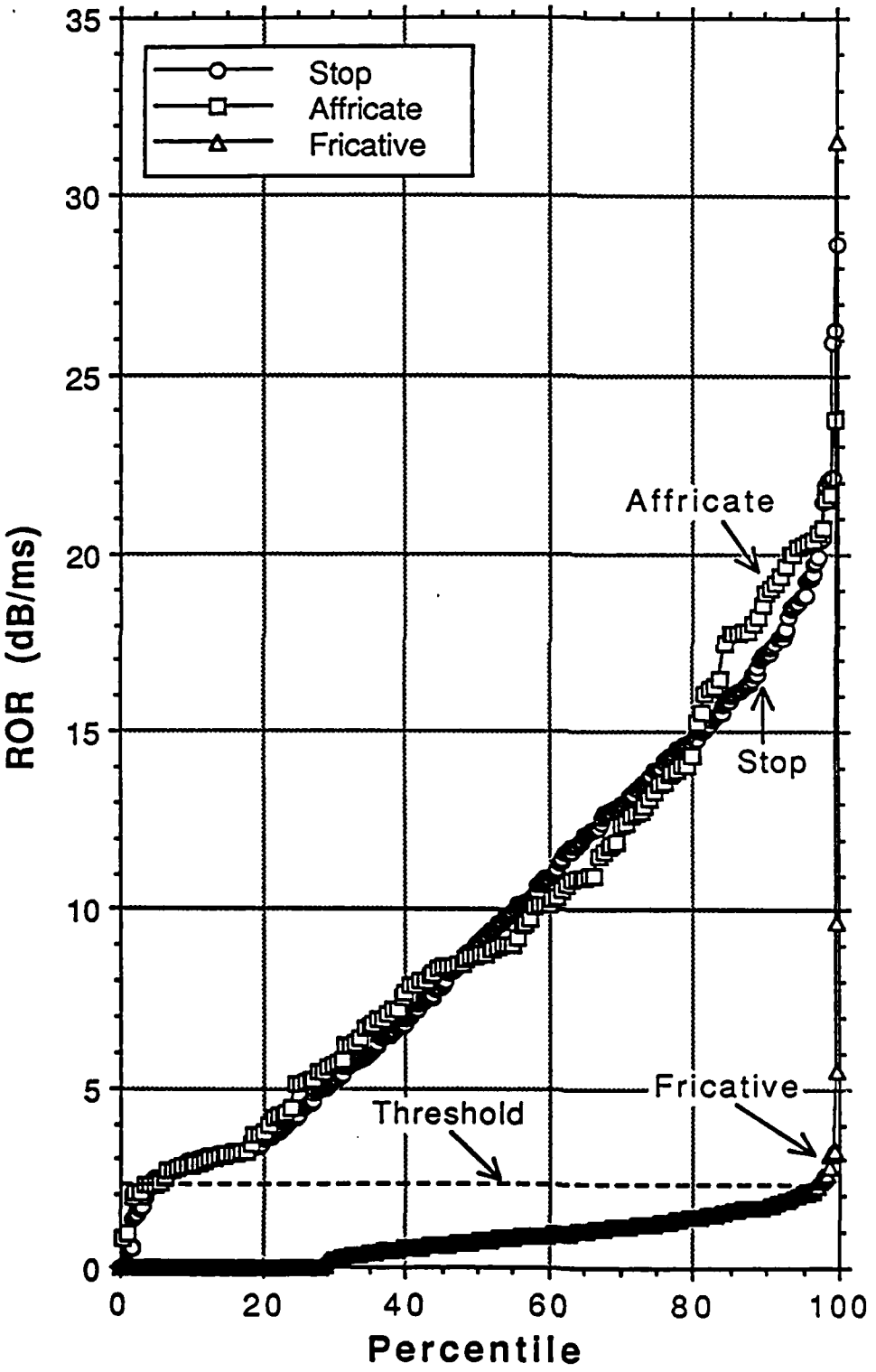

optimal CF
CF=0 Hz
100
95
90
85
80
75
70
Percent Correct
100
1000
10000
Cutoff Frequency (Hz)

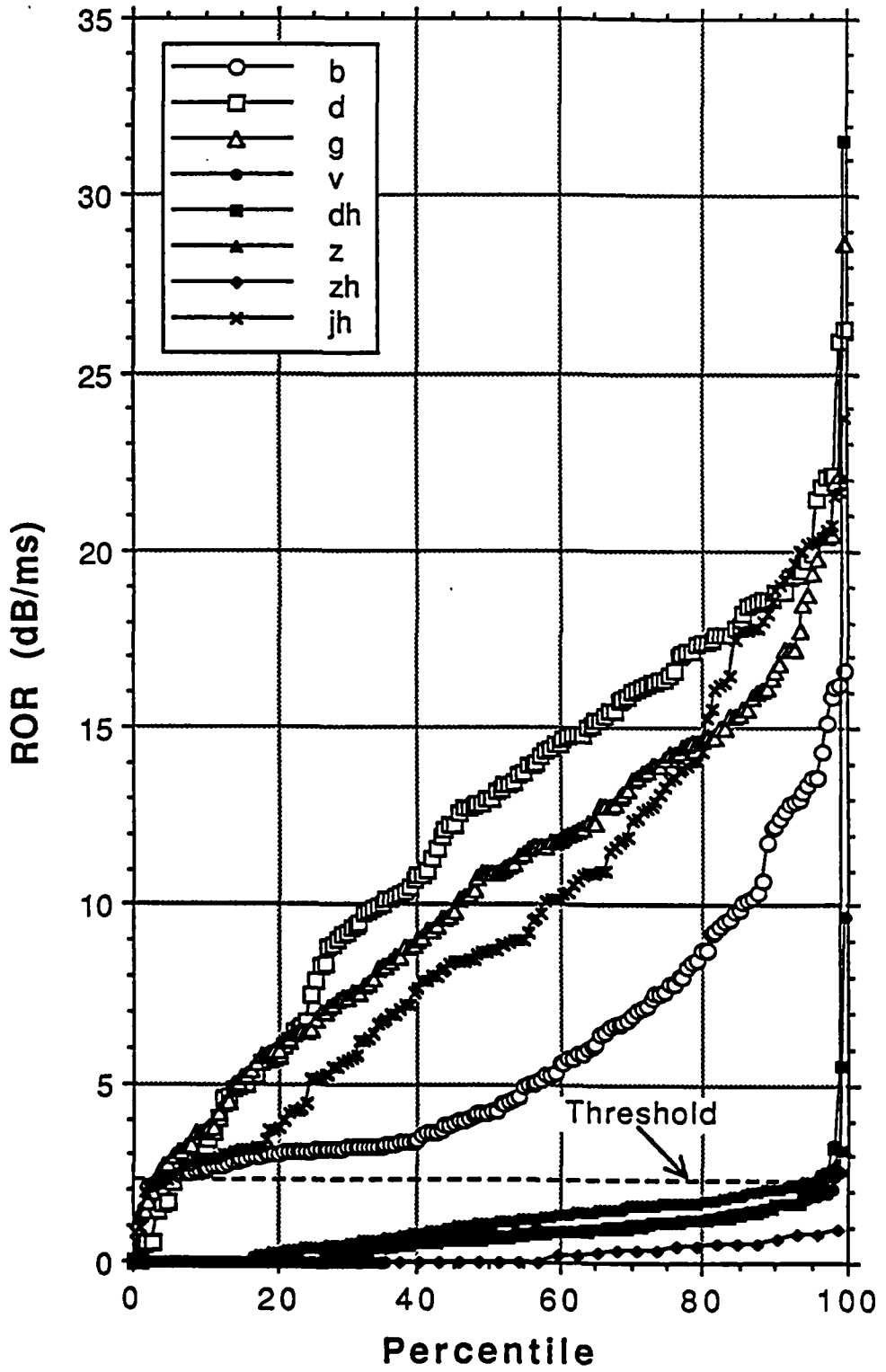

GUT.
16384
Waveform
Amplitude
0
-16384
60
50
40
30
20
10
0
RMS
(dB)
15.0
10.0
5.0
0.0
-5.0
ROR
(dB/ms)
10000
5000
0
ZC
(#/s)
0.100
0.200
0.300
0.400
0.500
Time (s)

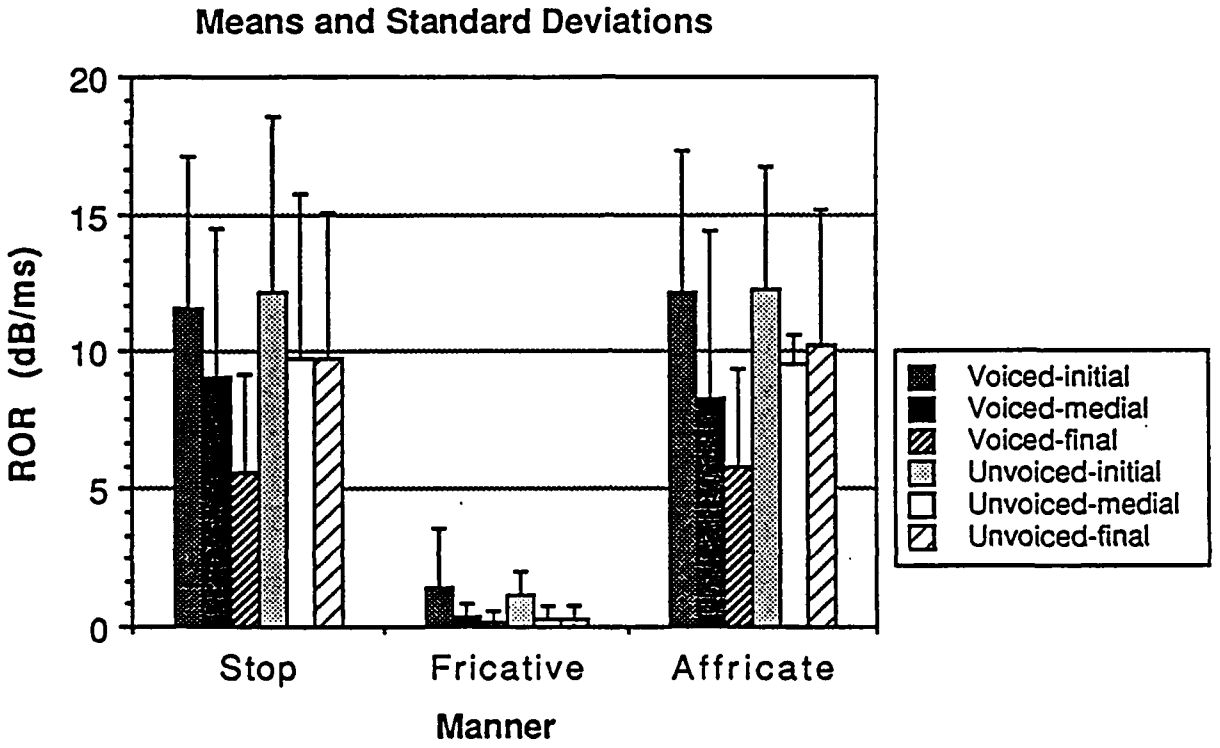

VET.
16384
Waveform
Amplitude
0
-16384
60
50
40
30
20
10
0
RMS
(dB)
2.0
ROR
(dB/ms)
0.0
-2.0
10000
ZC
(#/s)
5000
0
0.100
0.200
0.300
0.400
0.500
Time (s)


BOOTH.
16384
Waveform
Amplitude
0
-16384
60
50
40
RMS
(dB)
30
20
10
0
2.0
ROR
(dB/ms)
0.0
-2.0
10000
ZC
(#/s)
5000
0
0.100
0.200
0.300
0.400
0.500
Time (s)


ZOOT.
16384
Waveform
Amplitude
0
-16384
60
50
40
RMS
(dB)
30
20
10
0
2.0
ROR
(dB/ms)
0.0
-2.0
10000
ZC
(#/s)
5000
0
0.100
0.200
0.300
0.400
0.500
Time (s)


Stop
Affricate
Fricative
ROR (dB/ms)
Percentile
Affricate
Stop
Fricative
Threshold
0
5
10
15
20
25
30
35
0
20
40
60
80
100


b
d
g
v
dh
z
zh
jh
35
30
25
20
15
10
5
0
ROR (dB/ms)
Threshold
0
20
40
60
80
100
Percentile


Means and Standard Deviations
ROR (dB/ms)
20
15
10
5
0
Stop
Fricative
Affricate
Manner
Voiced-initial
Voiced-medial
Voiced-final
Unvoiced-initial
Unvoiced-medial
Unvoiced-final

References

Borden, G. J. and Harris, K. S. (1984). *Speech Science Primer.* Williams and Wilkins, Baltimore, MD.

Carterette, E. C. and Jones, M. H. (August 1974). On the statistics of spoken american english. *Proceedings Speech Communication Seminar, Stockholm*, 3:165–173.

Gerstman, L. J. (1957). *Perceptual Dimensions for the Friction Portions of Certain Speech Sounds.* PhD thesis, New York University.

Halle, M., Hughes, G. W., and Radley, J. (1957). Acoustic properties of stop consonants. *Journal of the Acoustical Society of America*, 29:107–116.

Howell, P. and Rosen, S. (1983). Production and perception of rise time in the voiceless affricate/fricative distinction. *Journal of the Acoustical Society of America*, 73:976–984.

Hughes, G. W. and Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, 28:303–310.

Klatt, D. H. (1975). Voice onset time, frication and aspiration in word-initial consonant clusters. *Journal of Sp. and Hearing Research*, 18:686–706.

Klatt, D. H. (1979). Synthesis by rule of segmental durations in english sentences. In Lindblom, B. and Ohman, S., editors, *Frontiers of Speech Communication Research*, pages 287–300. Academic Press, New York.

Kuhl, P. K. and Miller, J. D. (October 3, 1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, 190:69–72.

Lamel, L. F., Rabiner, L. R., Rosenberg, A. E., and Wilpon, J. G. (August 1981). An improved endpoint detector for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(4):777–785.

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustic measurements. *Word*, 20:384–422.

Malecot, A. (1966). Mechanical pressure as an index of 'force of articulation'. *Phonetica*, 14:169–180.

Malecot, A. (1970). The lenis-fortis opposition: Its physiological parameters. *Journal of the Acoustical Society of America*, 47:1588–1592.

Miller, J. D. (1987). Auditory-perceptual processing of speech waveforms. In Yost, W. A. and Watson, C. S., editors, *Auditory Processing of Complex Sounds*, pages 257–266. Lawrence Erlbaum Associates, Hillsdale, NJ.

Miller, J. D. (May 1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, 85(5):2114–2134.

Oppenheim, A. V. and Schafer, R. W. (1975). *Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ.

O'Shaughnessy, D. (1987). *Speech Communication Human and Machine.* Addison-Wesley, Reading, MA.

Paliwal, K. K. (April 1984). Effect of preemphasis on vowel recognition performance. *Speech Communication*, 3(1):101–106.

Paliwal, K. K. and Rao, P. V. S. (1977). Acoustic phonetic recognition of continuous speech. *9th International Congress of Acoustics, Spain.*

Paliwal, K. K. and Rao, P. V. S. (1982). Synthesis-based recognition of continuous speech. *Journal of the Acoustical Society of America*, 71:1016–1024.

Pickett, J. M. (1980). *The Sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception.* University Park Press, Baltimore.

Rabiner, L. R. and Sambur, M. R. (February 1975). An algorithm for determining the endpoints of isolated utterances. *Bell System Technical Journal*, 54(2):297–315.

Reddy, D. R. (1966). Segmentation of speech sounds. *Journal of the Acoustical Society of America*, 40(2):307–312.

Regel, P. (1982). A module for acoustic-phonetic transcription of fluently spoken german speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-30:440–450.

Schwartz, R. and Makhoul, J. (February 1975). Where the phonemes are: Dealing with ambiguity in acoustic-phonetic recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):50–53.

Slis, I. H. (1971). Articulatory effort and its durational and electromyographic correlates. *Phonetica*, 23:171–188.

Stevens, K. N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, 68:836–842.

Stevens, K. N., Blumstein, S. E., and Glicksman, L. B. (1987). Voicing distinction for fricatives: Acoustic theory and measurements. *Journal of the Acoustical Society of America*, 82 Supplement 1.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, 3:32–49.

Upton, H. W. (March 1968). Wearable eyeglass speechreading aid. *Gallaudet Conference on Speech Aids - American Annals of the Deaf*, 113(2):222–229.

Weigelt, L. F., Sadoff, S. J., and Miller, J. D. (In Press). Plosive/fricative distinction: The voiceless case. *Journal of the Acoustical Society of America.*

Weinstein, C. J., McCandless, S. S., Mondshein, L. F., and Zue, V. W. (February 1975). A system for acoustic-phonetic analysis of continuous speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):54–67.

Tempo, stress, and vowel reduction in American English

Marios Fourakis

818 S. Euclid Ave.

Central Institute for the Deaf

St. Louis, MO 63110

Submitted to: The Journal of the Acoustical Society of America

Received:

AFOSR-Grant G-AFOSR-86-0335
Final Technical Report
Appendix

H

ABSTRACT

Two processes that affect the acoustic characteristics of vowels are discussed, phonological and phonetic vowel reduction. Phonological vowel reduction applies to unstressed vowels. Phonetic vowel reduction is supposed to apply to all vowels and be caused by fast speech rates, context, as well as lack of stress. In this experiment, the effects of changes in stress and in rate of speech (tempo) on the acoustic characteristics of American English monophthongal, nonretroflex vowels were examined. Four male and four female native speakers produced these vowels in two contexts, [h_d] and [b_d], in a carrier sentence, under four conditions of tempo-stress (slow-stressed, slow-unstressed, fast-stressed, and fast-unstressed). A total of 2304 vowel utterances were collected (8 speakers x 9 vowels x 4 repetitions x 2 contexts x 4 tempo-stress conditions). Measurements of duration and F0 showed that the subjects did in fact vary tempo and stress as instructed. The effect of a change in stress on vowel duration was found to be slightly larger than that of a change in tempo. The putative vowel portion of each utterance was analyzed, formant tracks were obtained, and these were plotted in an auditory-perceptual space [Miller, J. Acoust. Soc. Am. 85, 2114-2134 (1989)]. These plots served to determine the part of the utterance that could, in most cases, be considered its steady state, and could be represented by a point in the space, the coordinates of which were given by the average, over time, of the coordinates of the steady-state portion. The distance of these data points from the point representing the acoustic characteristics of a vowel produced by a neutral vocal tract was used to determine the magnitude of phonetic vowel reduction caused by faster tempo and less stress, relative to the slow-stressed condition. The results indicate that these distances do not have a major dependence on tempo and

stress. In addition, several vowel classifications schemes were tested using linear discriminant analysis, and the one proposed by Miller (1989) performed better than combinations of F0, F1, F2, and F3.

PACS numbers: 43.70Fq, 43.70Hs

INTRODUCTION

The term "vowel reduction" has two fundamentally different meanings, depending on whether it is used by a phonologist or a phonetician. On one hand, the phonologist uses it to refer to the phonological process whose application causes unstressed vowels to be realized as schwas. This process is evidenced in alternations, such as exhibited by the morpheme /tɛlɛ/ in the words "telegraphic" and "telegraphy." In "telegraphy," the second vowel, bearing main word stress, is realized as a front, mid, lax vowel [ɛ]. In "telegraphic," this vowel is realized as a schwa, the result of the application of the phonological process of vowel reduction. This process does not apply to all unstressed vowels of American English. Exceptions include [ɔ], [ʉ], and [ʊ] for monophthongal vowels, and [ɔɣ] and [ɑw] for diphthongal vowels (Ladefoged, 1982: p. 79). The process is independent of tempo and dependent only on the stress value assigned to a particular vowel. On the other hand, there are the phonetician's definitions of vowel reduction. The most common definition of vowel reduction is given by J. L. Miller (1981, p. 42): "Vowel reduction refers to the tendency for the obtained formant frequencies of a vowel to fall short of the idealized target values for that vowel - those values that would be obtained if the vowel were produced in isolation - resulting in an overall shrinkage of the vowel space." Consonantal context, destressing, and rate of speech (tempo) are listed as factors contributing to this kind of vowel reduction, which will henceforth be called phonetic vowel reduction to distinguish it from phonological vowel reduction. Another difference between phonetic and phonological vowel reduction is that there are no vowels exempt from

phonetic vowel reduction, in contradistinction to the exceptions of the phonological process.

In an important paper, Lindblom (1963) reviewed the literature on phonetic vowel reduction, which, in agreement with J. L. Miller (1981), is in almost all cases equated with vowel neutralization or vowel centralization, that is, a movement toward a schwa-like formant pattern. At the conclusion of his work, Lindblom himself sharply distinguishes formant undershoot (in his terms "reduction") from centralization or neutralization. Rather, he presents his "assimilation" theory, which, as explained below, may predict a movement away from a schwa-like pattern due to formant undershoot (Lindblom's reduction). In the present paper, the terms vowel reduction, vowel neutralization, and vowel centralization will be treated as synonymous, while Lindblom's vowel reduction by formant undershoot will be referred to as formant undershoot.

The experiment reported here aimed to determine the extent to which phonetic vowel reduction, brought about by destressing and increased tempo, affects the nine monophthongal, nonretroflex vowels of American English.

A. Background

Changes in stress and tempo can have two kinds of effects on sounds. These are effects on the temporal characteristics and effects on the spectral characteristics. Thus, the rest of this report is organized around these two kinds of effects.

Several factors that influence vowel durations have been determined, including segmental identity, segmental environment, syllable structure and word structure (for a review see, among others, Klatt, 1976 for American

English; Lindblom, Lyberg, and Holmgren, 1982 for Swedish; Lehiste, 1970; 1972). However, there are conflicting reports concerning the effects of the specific factors of tempo and stress. While there is general agreement that vowels tend to be shorter when they belong to a relatively unstressed syllable and also when spoken at a faster tempo, there is disagreement as to whether the effects of stress and tempo are equivalent, i.e., when either stress or tempo is varied, with the other held constant, which variable has a greater effect on vowel durations.

For example, Gay (1978) examined the effects of changes in stress and tempo on American vowels, and, for his speakers, a change in tempo affected vowel durations more than a change in stress affected them. These results were corroborated by Tuller, Harris, and Kelso (1982), who reported syllable durations in English under different conditions of stress and tempo. Both of their speakers showed greater effects of tempo than of stress, especially their second speaker. On the other hand, Fourakis (1986) reported equivalent effects of changes in stress and tempo on Greek vowel durations, showing a 25% reduction in vowel durations going from stressed to unstressed or from slow to fast tempo. More recently, Crystal and House (1988a, 1988b) reported durational measurements of all American English vowels, monophthongal and diphthongal, produced by slow and fast speakers, and found, contrary to previous results, that changes in stress conditions had a greater effect on the duration of vowels than did the different speaking rates. Thus, although destressing and faster tempo bring about reductions in vowel durations, it is not clear which has the greater effect.

One explanation for the differences in these results might be the use of different materials. Gay (1978) and Tuller et al. (1982) used nonsense syllables of the form [pipip], etc. Fourakis (1986) used mostly nonsense

words, which were much more similar to possible real words than the above. Finally, House and Crystal (1988a) made their measurements on connected text, read by two different groups of speakers, one classified as slow, the other as fast. In the present experiment, an attempt was made to reach a middle ground, in terms of naturalness, by constructing lists of real words, inserting them in a carrier sentence, and manipulating stress through the use of a dummy syllable.

Similarly, when the effects of tempo and stress on the spectral characteristics of vowels are examined, there are also conflicting reports on the nature of these effects, and their relationship to the temporal reductions brought about by these two factors. Lindblom (1963) examined eight Swedish vowels, stressed and unstressed, in slow and fast speech. He found that fast tempo and lack of stress resulted in a shortening of these vowels and that this shortening was strongly correlated with formant undershoot, defined as failure to reach target formant frequencies. This failure was furthermore dependent on consonantal context and did not necessarily imply vowel centralization. However, Lindblom (1963) has frequently been interpreted as stating that "as the syllable (and hence, vowel nucleus) decreased in duration, the vowel became reduced, or more schwa-like in character" (Miller, 1981, p.43). Whereas formant undershoot might result in such "reduction", this is not a necessary consequence. For example, from Lindblom's Tables I and II, in the sequence [gæg], the vowel [æ] would have the target F2 frequency of 1625 Hz. The frequency of the second formant, after the release of [g] would decrease from its initial value of 2100 Hz, and, given enough time, it would reach the target. In a very short vowel, it would not reach the target of 1625, and therefore be further away from the typical schwa F2 frequency of 1500. Thus, in this

case, F2 would be shifted away from the schwa and even though there would be formant undershoot, this would be the opposite of centralization. However, for the vowel [θ] (in Lindblom's notation), in the same context, failure to reach the second formant target frequency from a preceding [g] would bring the second formant frequency closer to a centralized vowel. Therefore, the extent of phonetic vowel reduction due to shorter durations would be context and vowel dependent, and thus, as noted above, Lindblom, himself, argued for "assimilation", rather than centralization or neutralization.

Despite this caveat, research in the years following Lindblom's article, has concentrated on establishing whether shorter vowel durations brought about by increased tempo correlate with vowel reduction and shrinkage of the vowel space. The results have not always been consistent. Gay (1968) reported such a correlation for American English diphthongs, but when Gay (1978) measured durational and formant values for American English [i] and [ɑ] in CVC-type nonsense syllables, under different conditions of stress and tempo, he did not find these correlations. More recently, Bernstein-Ratner (1985), in examining the speech of mothers directed towards their children or towards adults, found that although vocalic durations were shorter in the adult-directed utterances, there was little or no correlation between formant frequencies and durations for American English vowels.

Overall, while destressing and faster tempo bring about decreases in vowel durations, the question of whether and which vowels undergo phonetic vowel reduction due to tempo and stress, has not been conclusively answered. The experiment reported here is an attempt to provide an answer to this question.

B. Theoretical framework

The possibility of vowel reduction during the production of vowels in different contexts and under different conditions of stress and tempo presents a serious problem to theories of phonetic recognition that employ bottom-up procedures in deriving phonetic representations from the incoming acoustic signal. Theories like those proposed by Miller (1989) or Stevens and Blumstein (1979, 1981) are crucially dependent on the premise that there is information in the speech signal that is invariant under common transformations such as increases of tempo or reduction of stress. For example, Miller's (1989) auditory-perceptual theory of vowel recognition utilizes the first three significant prominences in the short-term spectrum of a vowel, as well as the fundamental frequency during the vowel's production, in order to define a space in which productions of like vowels can be grouped together. This is accomplished through the use of the following three equations:

$$x = \log (SF3/SF2) \qquad \text{Eq. 1}$$

$$y = \log (SF1/SR) \qquad \text{Eq. 2}$$

$$z = \log (SF2/SF1) \qquad \text{Eq. 3}$$

where SF1, SF2, and SF3 are the first three significant prominences in the short term spectrum of a vowel utterance, and SR is a sensory reference dependent on the current speaker's fundamental frequency and is defined by Equation 4:

$$SR = 168(GMF0/168)^{1/3} \qquad \text{Eq. 4}$$

where GMF0 is the geometric mean of the current talker's fundamental frequency.

Through the use of Equations 1-4, the short-term spectrum of any vowel production can be represented as a point in the three dimensional space defined by the x, y, and z coordinates. The position of the point is defined by the values of x, y, and z for that vowel production.

The transformation of vowel spectra through Equations 1-4 proved to be an efficient way of normalizing vowel productions across speakers of different sex and age, and across some contexts (Miller, 1989). However, given the putative "shrinkage" of vowel space suggested by J. L. Miller (1981) when changes in tempo and stress are introduced, it is not clear how the concept of target zones, and of the three-dimensional space defined by the x, y, and z coordinates, could account for changes, if any, in vowel spectra brought about by these changes. It also remains to be seen how this classification procedure compares with more traditional procedures, such as F1 by F2 classification.

I. METHOD

A. Subjects and speech material

The subjects were four female and four male native speakers of Midwestern American English, with no known speech or hearing disorders. They ranged in age from eighteen to sixty years. All were associated with Central Institute for the Deaf, either as research scientists or as research assistants.

Two lists of words were constructed, containing the nine, non-retroflex, monophthongal vowels of American English, in two contexts: i. [h__d], as was used by Peterson and Barney (1952), and ii. [b__d]. These words are shown in Table I. As mentioned in the Introduction, having real words as target syllables in a carrier sentence was considered a reasonable compromise between nonsense words and continuous read text. Thus, the two contexts were chosen because they were minimally different, and afforded seventeen real and one nonsense word as target syllables. The one nonsense word was matched by a real word of similar structure so that the subjects could compare the two and pronounce the correct vowel. Two carrier sentences, also shown in Table II, were constructed containing the dummy syllable "kay", which was included to create a two-syllable compound. In one carrier sentence this dummy syllable was written in lower case, and subjects were instructed to pronounce the two syllable compound with main stress on the second syllable, making it an iamb ("stressed" conditions). In the other carrier sentence, the dummy was in upper case, and the instruction was to pronounce the two-syllable compound with main stress on "KAY", thus making it a trochee ("unstressed" conditions). One of the male subjects could not perform this stress shifting task and was replaced. None of the other subjects reported having any difficulty with this task at the time the recordings were made. The different tempo conditions were obtained by instructing the subjects to speak at what they considered to be a rate appropriate for a lecture or talking to a non-native speaker (slow), and at a rate that was faster than the previous one, but with which they felt comfortable (fast). Six repetitions of each word were randomized and placed in the appropriate carrier sentence. These randomizations were different for each speaker and each tempo-stress condition. The first four

occurrences of each target syllable were used for the experiment. The other two were used only if some extraneous circumstance such as coughing, shifting of papers, etc. caused one of the first four to be discarded. This yielded 2304 utterances (2 (tempo) x 2 (stress) x 2 (context) x 9 (vowels) x 4 (repetitions) x 8 (subjects)).

B. Recordings

The speakers were recorded in an anechoic chamber using a low-noise microphone/pre-amplifier combination (Bruel & Kjaer 4179/2660). The microphone was placed at a height equal to the subject's mouth, at a distance of 1/2 meter (zero degrees angle of incidence). Talkers were instructed to speak with normal conversational effort, resulting in a signal of approximately 70 dBA at the microphone. The microphone output was channeled directly into a Sony PCM-501ES digital audio recorder (16 bit mode) with a JVC 720 VCR serving as the storage medium. The frequency response of the system ranges from 10 Hz to 10 kHz (+/- 2 dB) and the signal-to-noise ratio is greater than 65 dB A-weighted. The microphone is the limiting factor for these specifications. Subjects were instructed to begin reading tokens in order that a recording level in the -9 dB VU range could be selected. When that recording level had been set, a calibration tone was recorded for each subject. The calibration tone consisted of a 1 kHz sine wave generated by a Hewlett-Packard 3325A synthesizer at a constant output level of 69.5 mV (equivalent to 70 dB SPL at the microphone, 1V/Pa). Reference to this tone facilitates access to the original SPL of each subject's speech.

The recordings were digitized at 20 kHz with 16-bit precision using a MicroTechnology Digisound-16 analog-to-digital converter and interface, a 50-Hz analog high-pass filter (removing incidental low frequency noise), and a 10-kHz anti-aliasing filter. After digitization, files were digitally notch filtered at 60 Hz to remove any residual AC noise and stored on a MicroVAX II for later processing using the commercial software package ILS (Interactive Laboratory System).

C. **Measurements - durational**

At the time of digitization the following intervals were measured to the nearest millisecond, by hand from the waveform, as displayed on an HP2623 graphics terminal: the total sentence duration, the total target-syllable duration, and the duration of vocalic nucleus of the target-syllable. The total sentence duration was measured from the first glottal pulse indicating the onset of voicing for the word "I" of the carrier sentence to the last trace of the nasal murmur for the end of the word "again". The word "I" was often preceded by one or two glottal stops, which were not included in the total sentence duration. The target-syllable was then isolated from the carrier sentence, and an example is shown in Figure 1, which displays the waveform for the word "bad" as spoken by male subject number 1, in the slow-stressed condition. The vertical lines show the points taken as the onset and offset of the syllable and the vocalic nucleus. (Almost all the syllable final [d]'s were flapped, regardless of tempo or stress conditions.)

D. Measurements - spectral

After digitization, the waveforms were edited so that all but the part that was previously determined to correspond to the vocalic nucleus was set to zero. They were then analyzed using ILS, and linear prediction coding (LPC) and cepstral analysis to determine the fundamental frequency were performed. A 24 ms Hamming window moving in 1 ms steps was utilized for the analysis, with 24 poles, and a pre-emphasis factor of 98%. A set of fundamental and formant frequency values for each ms of waveform were extracted and were then stored in table-format file suitable for hand-editing. The algorithms of ILS frequently mistracked the formants or the fundamental frequency. In these instances, F0 was estimated by determining the time interval for three pitch pulses of the waveform, calculating the average period, and inserting the inverse of this value into the appropriate place in the table-format file. Where any one of the formants was incorrectly tracked, direct FFT's were performed on the waveform, or the root solving command of ILS was used to determine the formant values, which then were inserted into the appropriate places. After hand-editing was completed, each file was smoothed using a first-order resonator with a cut-off frequency of 20 Hz. Then, each set of F0, F1, F2, and F3 for each ms of waveform was converted to a set of x, y, z coordinates using Equations 1-4 above. Each set of coordinates defines a point in the three-dimensional APS, and a sequence of points defines a path through the space, corresponding to the original waveform. The distance between any two points is then a measure of spectral change over the period of time elapsed between the two points. Thus, one can determine the portion of the path that corresponds to a relative steady state in the original waveform by the

clustering of points close to one another over an interval of time. Using algorithms developed in-house, we then proceeded to take an average of the x,y,z values for this interval, as well as the geometric mean of the actual F0, F1, F2, and F3 from the formant file. The average of these xyz values was then used to represent this point in the APS. The set of F0 and formant values was used for comparison with the x,y,z coordinates in the several analyses reported in the Results section. All 2304 vowel utterances were submitted to this procedure, and steady-state intervals were determined for almost 85% of the utterances. For the other 15%, most of which were found in the fast-unstressed condition and comprised productions of lax vowels, no steady state could be determined. In these cases, the choice of any one point or set of points to represent the waveform would have been arbitrary, so it was decided to compute average x,y,z, and F0, F1, F2, and F3 over the entire vowel segment.

II. VERIFICATION OF CHANGES IN TEMPO AND STRESS

This section is divided into two parts, which discuss, in turn, the effects of changes in tempo and stress on durations (sentence, syllable, vocalic nucleus) and on the fundamental frequency of the part of the vowel utterance chosen to represent the utterance. The section on fundamental frequency is included as further evidence that the subjects shifted stress away from the target syllable, since it has been shown that higher F0 is a correlate of stressed vowels (Lieberman (1960). Furthermore, the values of F0 are necessary in order to transform any set of formant frequencies into a point in APS. All analyses of variance used the subjects as replicates, and were all 2 (tempo) x 2 (stress) x 2 (context) x 9 (vowel) factorials. F-

ratios and degrees of freedom are reported only for significant effects ($p<.01$) or marginally significant effects ($.01<p<.05$). In the discussion of the results, for briefness of exposition, the term "vowel identity" is used instead of "identity of the intended vowel." Durational results for sentences and syllables broken down by vowel identity are not presented, although they are discussed, in order to keep the number of tables within reasonable limits.

A. Durations

Table IIA. lists the mean durations in ms for sentences pooled across vowels and context, for each condition of tempo-stress. Sentences were shortened by 28% going from slow to fast tempo and this effect was significant ($F(1,7)=101.06$, $p<.01$). This result indicates that the subjects did in fact change their rate of speech as instructed. However, there were no effects of stress and context. Sentences containing low target vowels ([ɔ, ɑ, æ]) were longer than sentences containing any of the other vowels ($F(8,56)=5.01$, $p<.01$). This reflects the inherently longer durations of low vowels as opposed to high vowels (cf. below on vowel durations). There was only one siginificant interaction, between tempo and stress ($F(1,7)=17.17$, $p<.01$). This was because the sentence durations were affected by stress in the slow tempo but not in the fast.

Table IIB lists the mean durations in ms of the target syllables pooled across vowels, for each condition of tempo and stress. The change from slow to fast tempo shortened the target syllable by 29.5%, about the same as it shortened sentences, and this was significant ($F(1,7)= 36.78$, $p<.01$). The

change from stressed to unstressed shortened the syllable duration by 34.5%, which was also significant($F(1,7)=60.38$, $p<.01$). The different contexts did not significantly affect syllable durations. Syllables containing low vowels were significantly longer than those containing mid and high vowels ($F(8,56)=58.12$, $p<.01$.. There were three significant interactions: between stress and tempo ($F(1,7)=16.53$, $p<.05$); between stress and vowel identity ($F(8,56)=3.14$, $p<.01$); and between context and vowel identity ($F(8,56)=3.11$, $p<.01$). The two significant interactions involving vowel identity are discussed more fully in light of the vowel duration results below. The interaction between stress and tempo is indicative of the tendency of syllables (and vowels, cf. below) to resist shortening beyond a certain point, a tendency which has been called the "incompressibility effect" (Klatt, 1973). Thus, there is a considerable difference in durations between the slow-stressed condition and the slow-unstressed or fast-stressed conditions. However, when both shortening factors are present, i.e., in the fast unstressed condition, the differences are not as great, mainly, because the syllable, like the vowel it contains, cannot be shortened much further.

Tables III and IV list the mean durations of each vowel in each tempo and stress condition, for each context separately. The overall means indicate an overall shortening effect when going from slow-stressed to fast-unstressed conditions. The change from slow to fast tempo shortened vowels by 29%, an effect which was significant ($F(1,7)=33.89$, $p<.01$). The change from stressed to unstressed shortened vowels by 33%, which was also significant ($F(1,7)=48.42$, $p<.01$). This shortening effect was similar for the two rates and contexts. In the [b_d] context, a change in stress in the slow condition shortened vowels by 35%, while the same change in the fast condition shortened vowels by only 29%. In the [h_d] context the

percentages are 37 for the slow, and 29 for the fast tempo. These differences are probably due to vowel durations for these speakers approaching their incompressibility limits as discussed in Klatt (1973). The interaction between tempo and stress was marginally significant ($F(1,7)=7.53$, $p<.05$), indicative of the fact that the shortening effect was not cumulative but diminished as vowels became shorter and shorter. Although vowels were in general shorter in the [h_d] context than in the [b_d] context, this was not a significant effect. Vowel identity had the expected highly significant effect of lower vowels being longer in all conditions and contexts than mid and high vowels ($F(8,56)=107.06$, $p<.01$). Furthermore, the vowel identity interacted significantly with all three conditions: with tempo ($F(8,56)=7.66$, $p<.01$); with stress ($F(8,56)=7.66$, $p<.01$); and with context ($F(8,56)=9.81$, $p<.01$). In general, vowels that were the longest in the slow-stressed [b_d] context were shortened more in any of the other conditions.

In order to assess the relative magnitude of the tempo and stress effects, and since context had neither a main effect nor did it interact with tempo and stress, vowel durations were pooled across contexts, and the mean differences were computed for each vowel in two separate cases: Stressed vs. unstressed, pooled across tempo; Slow vs. fast, pooled across stress. The resulting means and differences are shown in Table V. The shortening effect of stress is about 8 ms greater on the average than the shortening effect of tempo. Although the difference is small, a pairwise t-test for mean differences over vowels in both conditions showed that it was significant ($t(17)=.7.278$, $p<.01$). This is in agreement with Crystal and House (1988a, 1988b), but not with Gay (1978), who found the shortening effect of tempo to be 11 ms greater than that of stress.

B. Fundamental frequency

Table VI shows fundamental frequency means for men and women for each vowel in each stress and tempo condition, pooled across contexts. Certain well-established trends are evident. First, high vowels have higher F0's than lower vowels. This is regardless of subject sex or any other condition (cf. Shadle, 1985; Ladd and Silverman, 1984; for an opposing view, Umeda 1981). Second, all vowels have lower F0's when the subjects were supposed to shift stress away from them. This is another indication that the subjects followed instructions (for similar results regarding stress, cf. Cooper, Soares, Harn, and Damon, 1982). These trends were confirmed by analyses of variance, which were performed separately for men and women, again using subjects as replicates.

The results for women are discussed first. Stressed vowels had higher F0's, and the effect was marginally significant ($F(1,3)=25.58$, $p<.02$). There was a significant effect of vowel identity ($F(8,24)=16.19$, $p.<01$), with high vowels having higher F0's than low vowels. Vowels in the [h_d] context had slightly higher F0's than they did in the [b_d] context (3 Hz average difference), but, surprisingly, the effect was consistent and highly significant ($F(1,3)=144.44$, $p<.01$). Contrary to the results of Cooper et al. (1982), there was no significant effect of tempo. Only one significant interaction was found, between vowel and stress ($F(8,24)=6.79$, $p<.01$), with the high vowels showing a greater lowering of the fundamental frequency in the unstressed condition than the low vowels.

The results for men paralleled those for women, although less strongly. Stressed vowels had higher F0 than unstressed vowels and the effect was marginally significant ($F(1,3)=18.66$, $p<.03$). High vowels had higher F0

than low vowels, and this effect was highly significant ($F(8,24)=7.60$, $p<.01$). The difference between the two contexts was the same as for the women, but the effect was only marginally significant ($F(1,3)=17.62$, $p<.03$). There was also a marginally significant interaction between stress and vowel identity ($F(8,24)=2.40$, $p<.05$). Although not statistically significant, it should be noted that the men, unlike the women, had higher F0's in the fast tempo, regardless of the stress condition.

III. EVIDENCE RELATING TO PHONETIC VOWEL REDUCTION

The results, to this point, have shown that the subjects did increase their tempo, as demonstrated by the shortened durations of the intervals measured. They also shifted stress away from the target syllable, when instructed to do so, as both durational and F0 measurements have shown. Thus, the next step was to evaluate the effects, if any, of these two changes on the spectral characteristics of the vowels under examination.

Tables VII and VIII list the average frequencies of the first three formants in both conditions of tempo and stress, pooled across contexts, for men and women respectively. Although individual formant values do not seem to be affected by the different conditions, it is possible that the overall formant pattern of a vowel, as well as the whole vowel space, could be affected, since non-significant shifts of individual formant frequencies could add up to a significant shift of the total pattern. This possibility was evaluated using the three-dimensional space proposed by Miller (1989), in which any shrinkage of the vowel space can be thought of as a drift towards the position occupied in the space by a neutral vowel. First, the

position of such a point was calculated in this space. This point represents the spectral pattern of a neutral reference vowel, such as the one described in Fant (1970, p. 66), corresponding to uniform vocal tract, 17.6 cm long, with resonances at 500, 1500, and 2500 Hz for F1, F2, and F3, and a male F0 of 133 Hz. Alternatively, in the space proposed by Miller, a point representing a female neutral tract with resonances at 600, 1800, and 3000 Hz for F1, F2, and F3 and a female F0 of 230 Hz, could be used. However, since the mapping into the space through equations 1-4 normalizes for sex differences, this point would have exactly the same coordinates as the point representing the male vocal tract. Then, the Euclidian distance between each point representing a measured vowel utterance in the experiment and this neutral reference point was calculated in this space. If the vowel space did in fact shrink, then these calculated distances, for each vowel, should decrease, as the formant patterns would approximate the pattern of the neutral reference vowel. The mean distance for each vowel in each tempo-stress condition is shown in Table IX. Figure 2, which is a two dimensional-projection of the vowel space in x', y' coordinates. shows the mean position of each vowel in reference to the neutral point, which is represented by a large circle, roughly in the middle of the space. Sets of four symbols are encircled to indicate separate vowels. As can be seen, there are small decreases in the mean distances from the neutral point due to increases in rate or decreases in stress. In order to evaluate quantitatively any change in distance brought about by changes in tempo and stress, analyses of variance were performed on the distances of the vowels in each condition, using the same statistical model as with durations. Pair-wise t-tests on the mean distances were also performed, pairing each tempo-stress condition with all of the others. Further, the correlation was

calculated between the individual durational measurements and the distances from the neutral point for each speaker's mean for each vowel in each tempo, stress, and context condition.

The results of these analyses were as follows: The analyses of variance presented a very complicated picture, mainly because of the interactions caused by the vowels [ɪ] and [ɛ], which, in contradistinction to the other seven vowels, not only did not reduce, as measured by the distance metric, but became more distinct in the faster tempo and unstressed conditions. In general, there were significant effects only for vowel identity ($F(8,56)=74.27$, $p<.01$). There were marginally significant effects of: 1. context ($f(1,7)=9.87$, $p<.02$), with vowels in the [h_d] context being further from the neutral point than in the [b_d] context; and 2. stress ($F(1,7)=5.76$, $p<.05$), with unstressed vowels being about .015 log units closer to the neutral point than stressed vowels. There was no significant main effect of tempo. There were four highly siginificant interactions: 1. between tempo and vowel ($F(8,56)=4.88$, $p<.01$), with [i, æ, ɑ, ʌ, ɔ, ʊ, u] showing slight reduction of distance in the faster tempo, and [ɛ, ɪ] showing slight increases in distance in the faster tempo; 2. between stress and vowel ($F(8,56)=3.44$, $p<.01$), with the same two vowels showing increased distances when bearing less stress; 3. between tempo and context ($F(1,7)=36.70$, $p<.01$) with vowels in the [h_d] context remaining unaffected, while those in the [b_d] context decreased their distance from the neutral point in the faster tempo, by about .015 log units; and 4. between context and vowel ($F(8,56)=5.74$, $p<.01$), with vowels again being further away in the [h_d] context than in the [b_d] context, with the exception of [ʊ], for which the reverse situation obtained. There were no other two-way or three-way significant interactions.

The pair-wise t-tests of distance means, the results of which are shown in Table X, showed no significant differences. The comparison of the two most extreme conditions, Slow-stressed and Fast-unstressed, was the one closest to significance. It should be noted from this Table, that the biggest mean difference between conditions is .027 log units, which represents approximately 2.5 difference limens for this space, as reported in Hawks (1990a, 1990b).

The correlation coefficient between durations and distances over all vowels and conditions was .255 (r-square=.065). Specific correlations were calculated for each vowel. These correlations ranged from -.332 for [ɛ] (r-square=.11) to .408 for [ʊ] (r-square=.166). The negative coefficient for [ɛ] is indicative of the fact that the distances for this vowel increased as it shortened. Thus both the overall and the individual vowel results show that, as measured by the distance metric, the overall spectral pattern was by and large not correlated with vowel durations.

These results show that, even though subjects increased their tempo and shifted stress away from the target syllable, they still produced vowels with largely unaffected formant patterns. This result agrees with Gay (1978) and Berstein-Ratner (1985). It is also remarkable, that there was no correlation between the temporal shortening, induced by faster tempo and shifting of stress, and the position of the overall formant pattern in the vowel space relative to the point representing a neutral vocal tract. Thus, phonetic vowel reduction, defined as a consistent and sizeable shrinkage of the vowel space, is not a necessary consequence of destressing and increased tempo in English.

IV. COMPARISONS OF VOWEL CLASSIFICATION SCHEMES

One of the aims of this experiment was to evaluate how well multi-dimensional vowel classification schemes of the sort proposed by Miller (1989) could distinguish vowels produced under different conditions of stress and tempo, as compared to more traditional F1 by F2 schemes. It was deemed useful to use a statistical technique to compare x,y,z based classification with that based on combinations of F0, F1, F2, and F3. One statistical method to compare different schemes of classification is through the use of linear discriminant analysis, using as grouping variables the variables proposed as significant by each scheme. Linear discriminant analysis has been applied to problems of vowel classification by several investigators, including, but not limited to, Assmann, Nearey, and Hogan (1982), Syrdal (1985), and Syrdal and Gopal (1986). This analysis calculates the group mean for each decision variable and then assigns each token to be classified to a group on the basis of an a posteriori probability of group membership. In the initial phase we split the tokens into four groups, one each for every combination of tempo and stress, such that each group was composed of 576 tokens (9 vowels x 64 tokens/vowel). Using the R (resubstitution) method of classification, we performed a set of linear discriminant analyses, the results of which are shown in Table XI. The use of the x, y, and z variables resulted in vowel classification with higher percentage correct than all combinations of F0, F1, F2, and F3 in all four conditions. It is also important to note that the a posteriori probabilities of correct classification are much higher for x,y,z than for any combination of fundamental and formant frequencies, in most cases. These results together indicate that the use of the three transformed values

to represent these vowel points outperforms any combination of fundamental and formant frequencies, in all conditions of tempo and stress.

In order to evaluate the performance of these schemes with the total data set, all four conditions were pooled, creating a set of 2304 vowel points in x,y,z coordinates and 2304 sets of F0, F1, F2, and F3. The results of the linear discriminant analyses run on these sets are shown on the right hand side of in Table XI. Again, it can be seen that use of x,y,z was superior to all other tested schemes, both in terms of percentages and a posteriori probability of correct classification.

A stricter test of the efficiency of these schemes in correctly classifying the vowel productions would be to train the procedure on one set of tokens, e.g., the slow-stressed, derive a classification matrix, and use that matrix to classify the tokens in the other three conditions. Results for the matrices that performed the best in the slow stressed condition, i.e., the x,y,z classification matrix and the F0,F1,F2,F3 matrix, are shown in Table XII. It can be seen that the x,y,z matrix outperformed the F0,F1,F2,F3 matrix in all three conditions, and the a posteriori probabilities of correct classification were also greater (by .073 on the average) for the x,y,z matrix. These results provide an objective evaluation of the relative efficiency of the classification schemes tested.

V. GENERAL DISCUSSION

The experiment reported here was motivated by conflicting reports in the literature of the effects of tempo and stress on some of the acoustic characteristics of vowels, specifically durations and spectral patterns. Another major question was whether phonetic vowel reduction is an automatic

consequence of the temporal shortening brought about by increases in tempo and decreases in stress. The experiment also aimed to evaluate the potential of multi-dimensional vowel space schemes such as that proposed by Miller, to classify vowels whose acoustic characteristics might have been affected by changes in tempo and stress.

Concerning the temporal shortening effects, it was found that a change in stress had a slightly greater effect on vowel durations than a change in tempo. This difference, although small, was statistically significant. It seems that the effect of a change in tempo is global, involving all sub-units of an utterance. On the other hand, the shifting of stress from one syllable of a compound to the other is a local effect, and it should be expected that individual segments should be affected more than they would be by a change in tempo. On the other hand, it must also be taken into consideration that the difference between tempo and stress effects is small, and very similar to that found by Gay (1978), which was 11ms, albeit in the opposite direction. Crystal and House (1988b) report that the effect of stress is about 60 ms across all vowels, a value very similar to the one reported here, whereas the effect of fast vs. slow talkers in their study was only 10 ms. In their first article concerning this set of data, Crystal and House (1982) reported that the difference between slow and fast talkers was about 10%, in terms of time devoted to speech segments. They also point out that this classification of speakers into slow and fast, is different from the classical rate experiment, in which speakers are asked to speed up relative to some pre-established normal rate (i.e., as was done in the experiment reported here). The speakers in the present experiment increased their rates by as much as 30%, which is one possible reason that their tempo effect is very similar to the stress effect. Thus, it appears that the

differences between the present results and those of Crystal and House (1988a, 1988b) might be due to the different definitions of the slow and fast conditions. However, it is interesting in light of this consideration that the stress effects in the two experiments would be so similar.

The present data also showed that phonetic vowel reduction, defined as a shift of the formant pattern of vowels towards a neutral vowel, is minimal in American English, and dependent on various factors other than tempo and stress. Specifically, tempo had no reducing effects and stress had only a marginally significant effect, whereas context was the most significant factor influencing the relative position of the formant pattern of a given vowel utterance, a result similar to that of Stevens and House (1963). This result is somewhat surprising, given that in the present experiment only two contexts were employed, both of which are supposed to interfere minimally with vowel articulation. The results concerning tempo agree with Engstrand (1988), who found no effects of tempo on the tense point vowels of Swedish ([i,a,u]). Engstrand also points out that context might be more important in determining the extent of phonetic vowel reduction than either tempo or stress. One additional finding of the present experiment was that although increases in tempo and decreases in stress shortened vowel durations significantly, there was no correlation between durations and distances from the neutral point. In fact, some vowels, even though shorter in duration, were further away from the neutral point in the faster or unstressed conditions than in the slow-stressed. This could be due to formant undershoot having a reverse effect, that is, resulting in formant values which are further away from the neutral point formants than the target, as was discussed in the Introduction in reference to Lindblom's (1963) results. Phonetic vowel reduction, when it does occur, and when it is associated with

shortened durations due to increases in tempo and shifting of stress, must be distinguished from formant undershoot.

In comparing possible classification schemes, linear discriminant analyses were used to evaluate the relative efficiency of the three-coordinate representation vs. that of spectral peaks and fundamental frequency. As a statistical technique, linear discriminant analysis can be considered a neutral test comparing differing approaches. As was seen, x,y,z classification outperformed all combinations of F0, F1, F2, and F3. It is interesting to note that F1 by F2 classification is always the worst and that classification by all formants and F0 is always the second best, after x,y,z. However, the best correct classification performance reached only 84.20% for the x,y,z coordinates in the slow stressed condition, while the worst was 68.57% for F1, F2 in the slow-unstressed condition (cf. Table XI). There are two possible explanations why this value does not approach the percent correct classification rates achieved by Miller (1987b) for the data from Peterson and Barney (1952) and his own lab (cf. Introduction): One is that the present set of data was collected with methods different from Peterson and Barney's. The other possible explanation lies in the fact that in about 15% of the utterances, as described in the Method section, it was necessary to compute means of the variables over the whole waveform and this may have introduced noise into the data, since it may be that in those very short waveforms, vowel perception is determined by the total pattern of formant change and not by any static representation of the formant pattern, a position very similar to that proposed by Strange (1989) or Nearey (1989). In further research, native speakers will be asked to classify the utterances collected in this experiment, and the confusion matrices will be compared to those produced by the linear discriminant

analyses. These classifications will also be used to evaluate Miller's proposed perceptual target zones for vowels (Miller 1987a; 1987b).

The results indicated that most vowels underwent minimal spectral changes due to increased tempo and decreased stress, with [ɑ] and [ɔ] being the ones affected the most. These vowels are in the process of merging in American English and are treated by some investigators as indistinguishable from each other (e.g. Strange 1989). Thus, it is expected that changes in tempo and stress should not present a major obstacle for the auditory-perceptual theory, or any theory crucially dependent on invariant acoustic information being present in the speech signal, as discussed in the Introduction.

VI. CONCLUSION

Phonetic vowel reduction is a process distinct from phonological vowel reduction, in terms of input, output, and the conditions under which each occurs. As was discussed in the Introduction, some vowels may be exempt from phonological vowel reduction, but no vowels are exempt from the possibility of phonetic reduction. The output of the phonological process, for American English, is most often a centralized schwa, and sometimes a central, high, non-round barred [i]. The output of the phonetic process may be a version of the vowel which is not as centralized as schwa, but is also not as well defined as the vowel in a carefully articulated production. Phonological vowel reduction takes place <u>regardless</u> of tempo or context. Although phonetic vowel reduction is supposed to take place <u>because</u> of tempo, stress, and context, the results presented here indicate that tempo

and stress are not major factors in determining the occurence and extent of phonetic vowel reduction in Midwestern American English. The results of context variation in the present study suggest that it plays a role in phonetic vowel reduction. Further studies of the effects of context, such as those of Huang (1989), will help determine the extent to which context contributes to this kind of phonetic reduction.

ACKNOWLEDGMENTS

Partial reports of the research reported here were presented at the Spring 1989 Meeting of the Acoustical Society of America and at the Winter 1989 Meeting of the Linguistic Society of America. The author wishes to thank Allard Jongman for help with the design of the experiment, Amy Schafer for help with the grueling process of recording and digitizing, Caroline Monahan for help with the analyses of variance, James D. Miller, Janet Weisenberger, John W. Hawks, and Michael Gottfried for comments on earlier versions, Terrance M. Nearey for supplying the Linear Discriminant Analysis program, Betty Tuller and two other anonymous reviewers for valuable comments and suggestions. The research was supported by NIDCD Grant RO1-DC00296 to Central Institute for the Deaf.

References

Assmann, P.F., Nearey, T.M., and Hogan, J.T. (1982). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. 71, 975-989.

Berstein-Ratner, N. (1985). "Dissociations between vowel durations and formant frequency characteristics," J. Speech Hear. Res. 28, 255-264.

Cooper, W.E., Soares, C., Ham, A., and Damon, K. (1982) "The influence of inter- and intra-speaker tempo on fundamental frequency and palatalization," J. Acoust. Soc. Am. 73, 1723 -1730.

Crystal, T.H., and House, A.S. (1988a). "Segmental durations in connected-speech signals: Current results," J. Acoust. Soc. Am. 83, 1553-1573.

Crystal, T.H., and House, A.S. (1988b). "Segmental durations in connected-speech signals: Syllabic stress," J. Acoust. Soc. Am. 83, 1574-1585.

Engstrand, O. (1988). "Articulatory correlates of stress and speaking rate in Swedish VCV utterances," J. Acoust. Soc. Am. 83, 1863-1875.

Fant, G. (1970). Acoustic Theory of Speech Production. (Mouton, The Hague).

Fourakis, M. (1986). "An acoustic study of the effects of tempo and stress on segmental intervals," Phonetica 43, 172-188.

Gay, T. (1968). "Effects of speaking rate on diphthong formant movements," J. Acoust. Soc. Am. 44, 1570-1573.

Gay, T. (1978). "Effects of speaking rate on vowel formant movements," J. Acoust. Soc. Am., 63 223-230.

Hawks, J.W. (1989a). Perceptual Aspects of a Three-dimensional Vowel Space. PhD Dissertation, Washington University, St. Louis, MO.

Hawks, J.W. (1989b). "Difference limens for synthetic vowel spectra," J. Acoust. Soc. Am. Suppl. 1 87, S158 (A).

Huang, C.B. (1989). "Effects of consonant context and lexical stress on vowel formant frequencies in continuous speech," J. Acoust. Soc. Am. Suppl. 1 86 S124 (A).

Klatt, D.H. (1973). "Interaction between two factors that influence vowel duration," J. Acoust. Soc. Am. 54, 1102-1104.

Klatt, D.H. (1976). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. 59, 1208-1221.

Ladd, D.R., and Silverman, K.E.A. (1984) "Vowel intrinsic pitch in connected speech," Phonetica 41, 31-40.

Ladefoged, P. (1982). A Course in Phonetics. 2nd Edition. (Harcourt, Brace, Jovanovitch, New York).

Lea, W.A. (1980). Trends in Speech Recognition (Prentice-Hall, Englewood Cliffs, NJ), p.127.

Lehiste, I. (1970). Suprasegmentals. (MIT Press, Cambridge, MA).

Lehiste, I. (1972). "Timing of utterances and linguistic boundaries," J. Acoust. Soc. Am. 51, 1228-1334.

Lieberman, P. (1960). "Some acoustic correlates of word stress in American English," J. Acoust. Soc. Am. 32, 451-454.

Lindblom, B. (1963). "Spectrographic study of vowel reduction," J. Acoust. Soc. Am. 35, 1773-1781.

Lindblom, B., Lyberg, B., and Holmgren, K. (1982). "Durational patterns of Swedish phonology: Do they reflect short-term motor memory processes?" (Indiana University Linguistics Club, Bloomington, IN).

Miller, J.D. (1987a). "Auditory-perceptual processing of speech waveforms," in Auditory Processing of Complex Sounds, edited by W.A. Yost and C.S. Watson (Laurence Erlbaum Associates, Hillsdale, NJ), pp. 257-266.

Miller, J.D. (1987b). "Classification of vowel productions by means of perceptual target zones: A response to Ladefoged and Studdert-Kennedy," J. Acoust. Soc. Am. Suppl. 1 82, S82 (A).

Miller, J.D. (1989). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. 85, 2114-2134.

Miller, J.L. (1981). "Effects of speaking rate on segmental distinctions," in Perspectives on the study of Speech, edited by P.D. Eimas and J.L. Miller (Laurence Erlbaum Associates, Hillsdale, NJ), pp. 39-74.

Nearey, T.M. (1989). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. 85, 2088-2113.

Peterson, G.E., and Barney, H.L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175-184

Shadle, C.H. (1985). "Intrinsic fundamental frequency of vowels in sentence context," J. Acoust. Soc. Am. 78, 1562-1567.

Stevens, K.N., and Blumstein, S.E. (1979). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. 64, 1358-1368.

Stevens, K.N. and Blumstein, S.E. (1981). "The search for invariant acoustic correlates of phonetic features," in Perspectives on the study of Speech, edited by P.D. Eimas and J.L. Miller (Laurence Erlbaum Associates, Hillsdale, NJ), pp. 39-74.

Stevens, K.N. and House, A.S. (1963). "Perturbation of vowel articulations by consonantal context: An acoustical study," J. Speech Hear. Res. 6, 111-127.

Strange, W. (1989). "Dynamic specification of coarticulated vowels spoken in sentence context," J. Acoust. Soc. Am. 85, 2135-2153.

Syrdal, A.K. (1985). "Aspects of the model of the auditory representation of American English vowels," Speech Communication 4, 121-135.

Syrdal, A.K., and Gopal, H.S. (1986). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. 79, 1086-1100.

Tuller, B., Harris, K.S., and Kelso, J.A.S. (1982). "Stress and rate: Differential transformations of articulation," J. Acoust. Soc. Am. 71, 1534-1543.

Umeda, N. (1981). "Influence of segmental factors on fundamental frequency in fluent speech," J. Acoust. Soc. Am. 70, 350-355.

Table I. Carrier sentences and target items

CARRIER SENTENCE:

I will say kay _____ again. (Stress on target item)

I will say KAY _____ again. (Stress on KAY)

TARGET ITEMS:

Context: B__D	H__D	IPA
bead	heed	[i]
bid	hid	[ɪ]
bed	head	[ɛ]
bad	had	[æ]
bahed	hod	[ɑ]
baud	hawed	[ɔ]
*bood	hood	[ʊ]
booed	*who'd*	[u]
bud	hud	[ʌ]

Four randomized lists of six repetitions of each word:

List 1: Slow rate - stress on target item

List 2: Slow rate - stress on KAY

List 3: Fast rate - stress on target item

List 4: Fast rate - stress on KAY

Table II. Mean durations and standard errors in ms pooled across contexts and vowels. N=576 for cell and 1152 for row and column means.

A. Sentences

	Stressed	Unstressed	Row means
Slow	1801	1595	1698
	s.e. 8.03	s.e. 7.62	s.e. 6.31
Fast	1218	1221	1220
	s.e. 6.17	s.e 9.25	s.e. 5.55
Column means	1510	1408	1459
	s.e. 9.98	s.e. 8.14	s.e. 6.53

B. Target syllables

	Stressed	Unstressed	Row means
Slow	387	237	312
	s.e. 2.74	s.e. 2.19	s.e 2.81
Fast	254	187	220
	s.e. 2.34	s.e. 1.50	s.e. 1.71
Column means	319	212	266
	s.e. 2.66	s.e. 1.52	s.e. 1.9

Table III. Mean durations and standard errors in ms for each vowel under each condition of stress and rate in context [b_d]. N=32 for cell and 288 for column means.

VOWEL	SS	SU	FS	FU
[i]	206	124	135	94
	s.e. 8.9	s.e. 4.8	s.e. 5.9	s.e 3.0
[ɪ]	153	96	95	71
	s.e. 7.5	s.e. 4.6	s.e. 4.3	s.e. 2.5
[ɛ]	168	115	119	91
	s.e. 6.5	s.e. 4.5	s.e. 4.1	s.e. 3.3
[æ]	251	170	175	127
	s.e. 8.5	s.e. 5.6	s.e. 5.7	s.e. 3.6
[ɑ]	261	174	186	134
	s.e. 7.1	s.e. 7.4	s.e. 5.0	s.e. 4.1
[ɔ]	254	166	180	127
	s.e. 8.1	s.e. 5.3	s.e. 6.6	s.e 3.5
[ʊ]	167	110	122	83
	s.e. 5.9	s.e. 4.8	s.e. 5.9	s.e. 2.8
[u]	213	137	149	104
	s.e. 9.6	s.e. 5.6	s.e. 6.5	s.e. 2.8
[ʌ]	171	111	124	87
	s.e. 6.1	s.e. 4.4	s.e. 5.3	s.e. 3.1
Mean	204	134	143	102
	s.e. 3.5	s.e. 2.4	s.e. 2.5	s.e. 1.6

KEY: SS = SLOW - STRESSED SU = SLOW - UNSTRESSED
FS = FAST - STRESSED FU = FAST - UNSTRESSED

Table IV. Mean durations and standard errors in ms for each vowel under each condition of stress and rate in context [h_d]. N=32 for cell and 288 for column means

VOWEL	SS	SU	FS	FU
[i]	203	118	127	86
	s.e. 7.9	s.e. 5.6	s.e. 3.7	s.e. 4.3
[ɪ]	153	92	96	70
	s.e. 7.2	s.e 4.2	s.e. 4.5	s.e. 3.7
[ɛ]	155	105	108	82
	s.e. 4.8	s.e. 4.2	s.e. 4.1	s.e. 4.3
[æ]	227	142	153	113
	s.e. 8.8	s.e. 6.5	s.e. 5.7	s.e. 5.4
[ɑ]	224	145	144	105
	s.e. 6.6	s.e. 4.3	s.e. 3.8	s.e. 5.3
[ɔ]	240	152	158	109
	s.e. 7.5	s.e. 4.6	s.e. 5.6	s.e. 4.5
[ʊ]	152	103	107	71
	s.e. 5.0	s.e. 3.2	s.e. 3.5	s.e. 4.2
[u]	205	124	133	89
	s.e. 7.8	s.e. 5.5	s.e. 4.3	s.e. 4.3
[ʌ]	150	96	104	78
	s.e. 5.4	s.e. 3.4	s.e. 3.2	s.e. 3.6
Mean	190	120	125	89
	s.e. 3.1	s.e. 2.0	s.e. 2.0	s.e.1.7

KEY: SS = SLOW - STRESSED SU = SLOW - UNSTRESSED
FS = FAST - STRESSED FU = FAST - UNSTRESSED

Table V. Mean durations in ms for each vowel for tempo across stress and stress across tempo

	SLOW	FAST	DIFFERENCE	STRESS	UNSTRESSED	DIFFERENCE
CONTEXT B_D						
[i]	165	115	50	171	109	62
[ɪ]	124	83	41	124	83	39
[ɛ]	142	105	37	144	103	41
[æ]	210	151	59	213	148	65
[ɑ]	218	160	58	224	154	70
[ɔ]	210	153	57	217	146	71
[ʊ]	140	102	38	145	97	48
[u]	175	126	49	181	120	61
[ʌ]	141	105	36	148	99	49
MEAN	169	122	47	174	118	56
CONTEXT H_D						
[i]	160	106	54	164	102	62
[ɪ]	123	83	40	125	82	43
[ɛ]	130	95	35	132	94	38
[æ]	184	133	51	190	127	63
[ɑ]	185	125	60	185	125	60
[ɔ]	196	134	62	199	131	66
[ʊ]	127	89	38	129	87	42
[u]	165	111	54	169	107	62
[ʌ]	123	91	32	127	87	40
MEAN	155	107	48	158	105	53
MEAN BOTH CONTEXTS	162	115	47	166	111	55

Table VI. Fundamental frequency means amd standard errors in Hz pooled across contexts

Condition / Vowel	Sex		SS	SU	FS	FU	Mean
[i]	M		133	100	150	97	120
		s.e.	3.2	1.5	6.4	4.7	
	F		233	188	239	189	212
		s.e.	3.5	5.4	2.4	4.2	
[ɪ]	M		130	98	147	103	120
		s.e.	3.4	1.2	5.6	1.9	
	F		233	184	235	188	210
		s.e.	2.3	5.5	2.8	5.0	
[ɛ]	M		124	94	146	99	115
		s.e.	2.7	1.4	5.9	2.0	
	F		219	177	223	181	200
		s.e.	3.1	5.3	2.3	3.9	
[æ]	M		116	93	139	97	110
		s.e.	2.1	2.7	5.0	1.8	
	F		214	173	213	176	194
		s.e	3.3	5.3	3.3	4.5	
[ɑ]	M		118	91	141	98	112
		s.e.	2.6	1.6	6.1	2.0	
	F		217	175	219	175	197
		s.e.	2.4	4.3	2.4	4.5	
[ɔ]	M		121	95	140	98	114
		s.e.	2.3	1.3	5.3	1.7	
	F		217	176	216	175	196
		s.e.	2.8	4.6	2.2	6.8	
[ʊ]	M		136	104	151	105	124
		s.e.	3.9	1.5	5.9	4.0	
	F		240	193	236	188	214
		s.e.	2.8	4.1	2.8	3.1	

Table VI continued

Condition Vowel	Sex		SS	SU	FS	FU	Mean
[u]	M		136	101	153	107	124
		s.e	3.3	1.4	6.1	1.7	
	F		244	189	243	191	217
		s.e	3.8	5.1	2.1	3.4	
[ʌ]	M		127	99	141	104	118
		s.e.	3.1	1.3	5.0	2.0	
	F		226	183	223	186	205
		s.e.	2.0	4.4	2.2	4.5	
Across all vowels	M		127	97	145	100	117
	F		227	182	227	183	205

SS = Slow - stressed
SU = Slow - unstressed
M = Male
FS = Fast - stressed
FU = Fast - unstressed
F = Female

Table VII. Mean formant frequencies and standard errors in Hz for the female subjects pooled across contexts. N=32 for cell means

Condition		Slow Stressed		Slow Unstressed		Fast Stressed		Fast Unstressed	
Vowel		Mean	s.e.	Mean	s.e.	Mean	s.e.	Mean	s.e
[i]	F1	328	5.5	344	8.0	341	7.2	354	9.4
	F2	2848	40.3	2752	34.1	2833	39.4	2719	38.1
	F3	3352	28.3	3332	40.0	3353	23.0	3204	50.6
[ɪ]	F1	468	0.5	453	5.8	468	0.5	450	6.5
	F2	2268	41.2	2211	31.5	2282	37.8	2234	36.8
	F3	3034	34.0	2944	29.0	3059	34.0	2952	32.6
[ɛ]	F1	635	8.8	621	15.4	630	13.0	590	13.2
	F2	2125	41.0	2059	34.7	2070	42.4	2073	42.8
	F3	3024	33.5	2906	28.9	2996	35.8	2898	38.7
[æ]	F1	752	20.1	729	13.4	775	16.3	702	13.8
	F2	2122	30.0	1953	21.9	1980	36.5	1978	21.1
	F3	2904	29.0	2808	40.3	2772	64.5	2728	40.8
[ɑ]	F1	961	35.5	1021	40.8	920	31.0	846	28.1
	F2	1325	18.5	1406	17.7	1301	13.0	1487	21.6
	F3	2832	57.2	2834	41.5	2844	60.5	2803	56.2
[ɔ]	F1	782	25.4	834	25.4	842	32.0	743	19.6
	F2	1106	22.8	1241	24.5	1147	22.5	1293	29.3
	F3	2769	65.4	2783	40.8	2790	45.4	2762	44.4
[ʊ]	F1	481	4.2	450	10.1	479	5.2	474	4.9
	F2	1314	33.2	1383	42.4	1418	47.0	1527	42.6
	F3	2890	37.4	2813	28.8	2869	34.8	2772	36.0
[u]	F1	382	10.9	377	12.0	432	7.9	370	10.7
	F2	1235	40.0	1401	39.8	1331	35.7	1372	39.3
	F3	2833	22.2	2787	25.2	2822	24.9	2811	24.6
[ʌ]	F1	676	10.6	667	13.7	665	8.4	646	15.2
	F2	1499	24.9	1617	22.3	1557	21.0	1677	27.3
	F3	2962	32.6	2815	32.0	2905	30.6	2820	35.1

Table VIII. Mean formant frequencies and standard errors in Hz for the male subjects pooled across contexts. N=32 for cell means.

Condition		Slow Stressed		Slow Unstressed		Fast Stressed		Fast Unstressed	
Vowel		Mean	s.e.	Mean	s.e.	Mean	s.e.	Mean	s.e.
[i]	F1	303	2.6	303	2.3	311	0.5	308	0.8
	F2	2379	34.7	2339	43.4	2406	36.4	2368	40.2
	F3	3162	66.0	2938	81.5	3052	68.3	2934	58.7
[ɪ]	F1	441	8.5	438	6.7	435	8.6	436	5.8
	F2	1917	38.0	1906	45.3	1940	47.3	1925	43.7
	F3	2705	39.1	2590	49.6	2682	44.8	2604	46.7
[ɛ]	F1	557	11.5	541	11.9	596	16.9	557	11.7
	F2	1783	36.2	1779	39.8	1795	38.1	1767	28.4
	F3	2642	42.8	2538	58.2	2607	41.4	2554	53.7
[æ]	F1	662	27.4	645	19.9	653	25.7	657	17.9
	F2	1739	22.0	1800	39.0	1775	21.7	1774	34.2
	F3	2555	43.3	2471	50.5	2525	41.9	2485	51.0
[ɑ]	F1	831	24.8	763	25.7	830	25.5	769	18.7
	F2	1184	11.4	1243	14.9	1222	14.6	1277	20.0
	F3	2616	30.0	2517	31.8	2550	45.1	2567	31.7
[ɔ]	F1	688	13.1	662	13.6	714	17.0	656	9.5
	F2	993	11.9	1059	11.0	1044	13.8	1092	13.7
	F3	2584	27.9	2492	26.7	2552	27.8	2547	30.8
[ʊ]	F1	443	8.9	444	8.4	474	6.7	452	4.7
	F2	1141	24.3	1224	20.3	1195	25.9	1276	23.1
	F3	2567	28.8	2542	39.5	2584	30.3	2525	39.3
[u]	F1	311	0.6	312	0.4	323	4.5	311	0.3
	F2	1018	18.0	1171	23.0	1087	21.4	1190	20.6
	F3	2440	35.5	2425	42.5	2450	28.3	2435	47.7
[ʌ]	F1	624	11.6	594	10.9	630	11.4	627	14.1
	F2	1285	13.1	1361	28.6	1298	20.5	1400	27.1
	F3	2624	28.0	2552	42.9	2587	31.4	2555	44.5

Table IX. Mean distances, in log units, of all points assigned to each vowel category from point representing a neutral vowel with F0=133 Hz, F1=500 Hz, F2=1500 Hz, and F3=2500 Hz.

Condition	Slow Stress	Slow Unstressed	Fast Stress	Fast Unstressed
[i]	.5149	.4894	.5152	.4854
[ɪ]	.2376	.2460	.2508	.2624
[ɛ]	.1035	.1240	.1184	.1279
[æ]	.1669	.1645	.1667	.1636
[ɑ]	.4090	.3880	.3842	.3307
[ɔ]	.3981	.3653	.3923	.3210
[ʊ]	.1892	.1679	.1747	.1442
[u]	.2665	.2437	.2375	.2481
[ʌ]	.2005	.1665	.1846	.1611
Mean	.2762	.2617	.2694	.2494

Table X. Statistical comparisons of mean distances from neutral point. SS = slow stressed; SU = slow unstressed; FS = fast stressed; FU = fast unstressed.

A. Pairwise t-tests.

Compared	DF	Mean X-Y	t	p(2-t)
SS - SU	8	.014	2.296	.0508
SS - FS	8	.007	1.328	.2208
SS - FU	8	.027	2.118	.0670
SU - FS	8	-.007	-1.773	.1142
SU - FU	8	.012	1.511	.1693
FS - FU	8	.020	2.019	.0781

Table XI. Percent correct classifications and a posteriori probabilities using linear discriminant analysis

Condition	Slow Stressed	Slow Unstressed	Fast Stressed	Fast Unstressed	Across conditions
Variables					
x,y,z	84.20	78.47	79.51	74.82	76.73
(APP)	(.74)	(.66)	(.68)	(.65)	(.66)
F1, F2	72.56	68.57	70.48	66.66	69.01
(APP)	(.57)	(.52)	(.54)	(.55)	(.53)
F0, F1, F2	77.95	73.95	71.85	72.56	72.65
(APP)	(.64)	(.57)	(.59)	(.61)	(.58)
F1, F2, F3	76.56	73.26	71.52	72.91	72.96
(APP)	(.61)	(.56)	(.58)	(.60)	(.57)
F0, F1, F2, F3	79.51	75.52	72.22	73.61	74.00
(APP)	(.65)	(.58)	(.60)	(.63)	(.59)

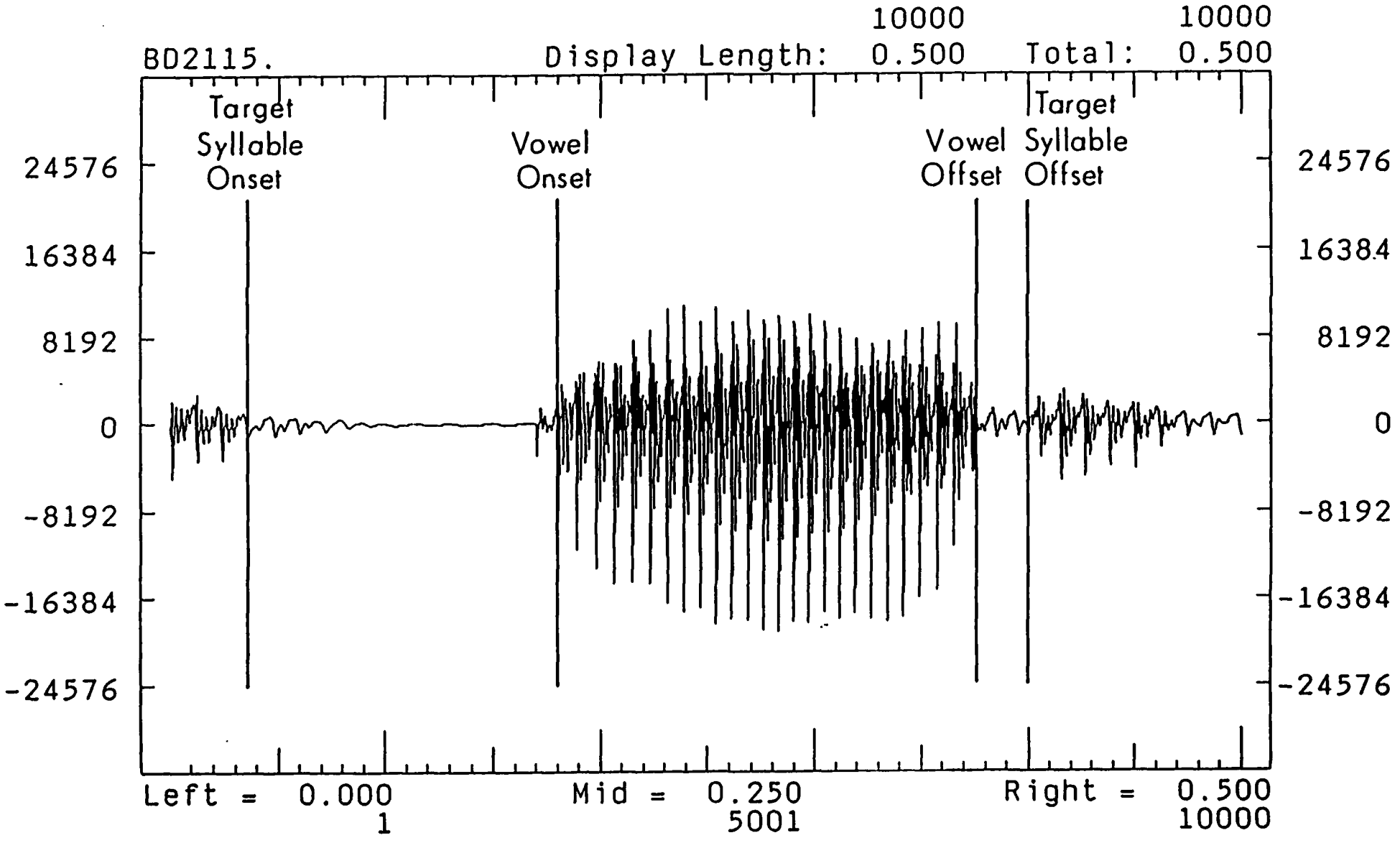
Table XII. Percent correct classification and a posteriori probabilities from linear discriminant analysis using the training matrices of the slow-stressed condition to classify the data of the other three. For the formant data, the matrix used was that of the F0,F1,F2,F3 classification.

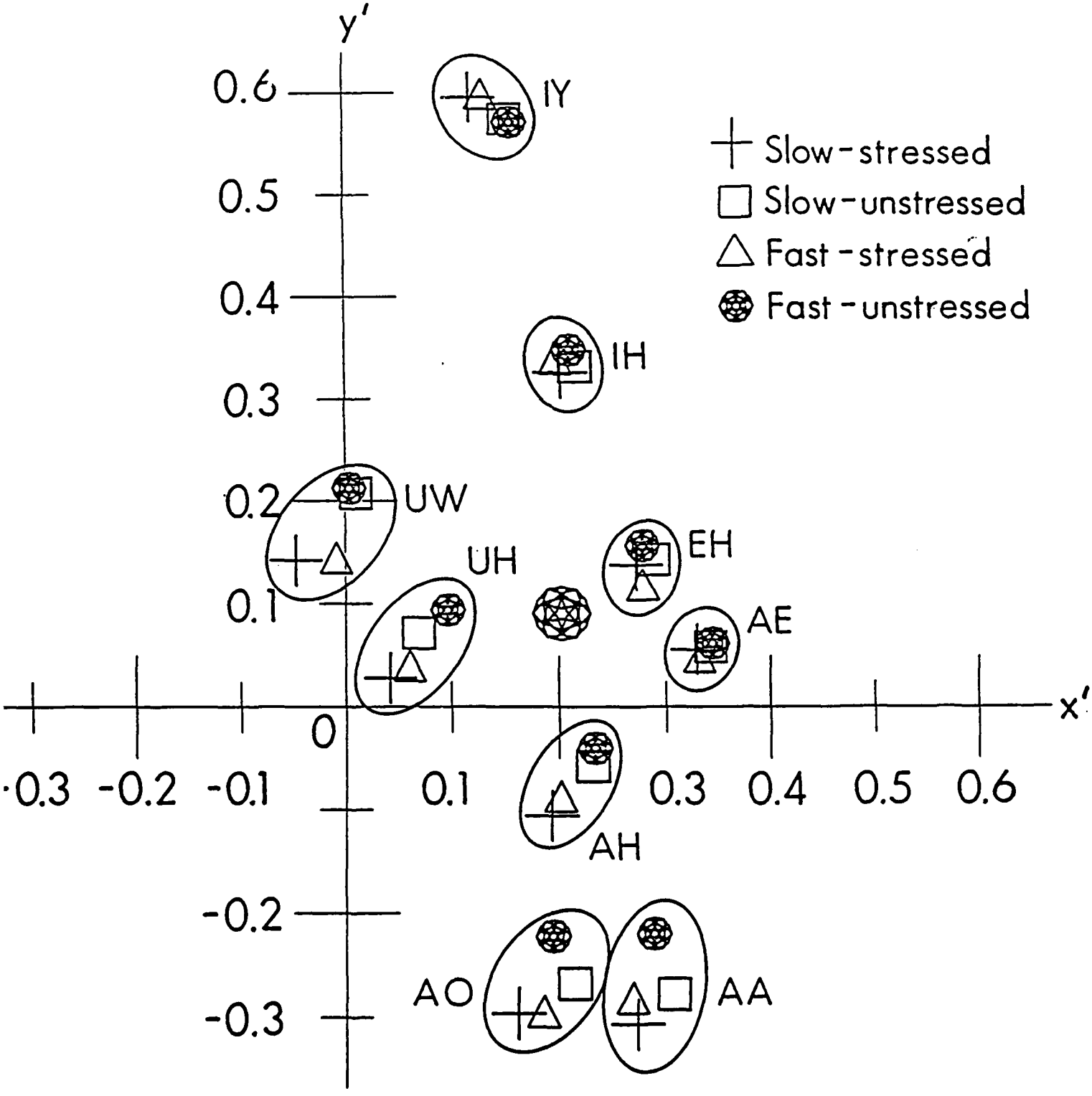
Data type	Slow Unstressed	Fast Stressed	Fast Unstressed
x,y,z	69.44	76.73	67.53
	(.63)	(.70)	(.60)
F0,F1,F2,F3	62.32	72.04	59.20
	(.56)	(.62)	(.53)

Figure Legends

Figure 1 The waveform of the target syllable "bad" spoken by Male subject 1, in the slow-stressed condition. The lines indicate, from left to right: i. the beginning of [b] (the onset of the target syllable), which starts with a voice bar, then is devoiced, has a very weak burst and friction component; ii. the onset of the vocalic nucleus; iii. the onset of the short syllable-final flap; and iv. the onset of the shwa of "again" (the offset of the target syllable).

Figure 2 Two-dimensional front view projection of the auditory-perceptual space, in x'y' coordinates, a transformations of the original APS coordinates as defined in Miller (1989). The axes are in 0.1 log units and the point of origin is (0, 0). The neutral point is represented by the large filled circle in the middle. The mean position of each vowel in each tempo-stress condition is plotted by a separate symbol. Sets of four symbols are encircled to indicate individual vowels. (IY=[i], IH=[ɪ], EH=[ɛ], AE=[æ], AA=[ɑ], AO=[ɔ], UH=[ʊ], UW=[u], AH=[ʌ])


BD2115.
Display Length: 10000 0.500
Total: 10000 0.500
Target Syllable Onset
Vowel Onset
Vowel Offset
Target Syllable Offset
24576
16384
8192
0
-8192
-16384
-24576
Left = 0.000 1
Mid = 0.250 5001
Right = 0.500 10000


y'
x'
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.2
-0.3
-0.3
-0.2
-0.1
0.1
0.3
0.4
0.5
0.6
IY
IH
UW
UH
EH
AE
AH
AO
AA
Slow-stressed
Slow-unstressed
Fast-stressed
Fast-unstressed

Mailed to Journal of Phonetics
1/17/91

An Approach to the Classification of American English Diphthongs

Michael Gottfried

and

James D. Miller

Central Institute for the Deaf

818 S. Euclid Avenue

Saint Louis, MO 63110

suggested running title: Classification of Diphthongs

AFOSR Grant G-AFOSR-86-0335
Final Technical Report
Appendix

I

Abstract

Six diphthongs of American English (/aU, aI, eI, oU, ɔI, ju/) were produced by four Midwestern American speakers (2 male, 2 female) at two tempos (slow, fast) with differing stress (stressed, unstressed) in two contexts ([b_d], [h_d]). Using a plot of the fundamental frequency and the first three formants derived from linear-prediction-coding (LPC) analysis, the onset and offset of each production was determined. The pattern of formants and fundamental frequency at the onset and offset of diphthongs was used to establish a set of parameters which can classify intended productions of the American English diphthongs in varying stress and tempo conditions with an average accuracy of 93 per cent. Results are also presented for diphthong, target-syllable, and sentence durations. The classification results are discussed with respect to hypotheses concerning the perception of diphthongs.

1. Introduction

Diphthongs, as a class of speech sounds, are commonly characterized in terms of movement from one vowel to another (Ladefoged, 1982, p.76). In phonetic transcriptions, diphthongs are accordingly rendered as a sequence of two monophthongal vowels (in Trager and Smith 1951 they are considered to consist of a vowel followed by a semivowel). Some investigators have sought to describe diphthongs with respect to sections isolated from a particular production which correspond to "vowel-targets". These targets are identified as sections of the analyzed signal in which the formants are essentially "steady-state", i.e., parallel to the time axis (Lehiste and Peterson 1961, Gay 1968). The intervening movement of one or more formants between the targets comprises a section termed a "glide". By this view, a diphthong is characterized as a sequence consisting of a steady-state followed by a glide and a terminal steady-state. If a diphthong production could be so segmented, one could then proceed to examine whether and in what way those segments delimited as "vowel segments" correspond to monophthongal vowels. Questions concerning the presence of steady-state segments in productions of American English diphthongs and their relation to monophthongs are addressed in sections 1.1 and 1.2, below.

Another view, which contrasts with that of investigators who have focussed on the relations between steady states in

diphthongs and monophthongal vowels, is to characterize diphthongs in terms of spectral and durational properties of the entire production of the diphthong. Some of these characterizations are described in sections 1.3 and 1.4.

1.1 Steady-state and glide segments

A number of studies have identified and characterized diphthongs with respect to their segmentation into steady-state and glide components. Lehiste and Peterson (1961) present a classification of /aU, aI, eI, oU, ɔI/ on this basis. For a production to be considered a diphthong they required that it exhibit "a vocalic nucleus containing two target positions" (p.277), each of which is associated with a steady-state section. On these grounds, they stated that /eI/ and /oU/ "should not properly be classed as diphthongs" (p.275) as the phoneme /eI/ was found to consist of a glide followed by a steady-state while /oU/ consisted of a typically short steady-state followed by a glide. The phonemes /aU ,aI, ɔI/ were, in contrast, considered diphthongs. However, the requirement that diphthong productions possess two steady-state target positions appears unduly restrictive, in particular when the effects of different speaking rates are considered. Gay (1968) examined the effects of differing rates of speech (slow, moderate, fast) on productions of /aU, aI, eI, oU, ɔI/. For slow and moderate speaking rates the presence of two steady-states was noted for <u>all</u> the phonemes studied but for the fast speaking rate

either the initial or final steady-state was found to be "negligible or not present".

1.2 Diphthongs and Monophthongs

The practice of transcribing diphthongs as a sequence of two monophthongs suggests that the diphthongs may in part be characterized in terms of formant values which identify those monophthongal segments. If particular diphthongs were identifiable with a distinctive set of initial and final monophthongal segments, such a characterization might provide a basis for distinguishing among the various diphthongs. Several studies have compared formant values at the beginning and end of diphthongs with those of the monophthongs by which the diphthong is usually transcribed. Such comparisons for the initial portion of the diphthong are presented in Table I; similar comparisons for the final portion are given in Table II. These studies differed in the manner by which these comparisons were established. Lehiste (1964), citing formant values (F_1, F_2 & F_3) reported in the earlier study with Peterson, compared average formant values for initial and final steady states of diphthongs with those obtained for productions of monophthongs by one of the subjects. Holbrook and Fairbanks (1962) based their comparisons upon formant values obtained at the beginning and end of various diphthong productions and those for a set of representative vowel samples as displayed in an F_1 by F_2 space. Wise (1965) compared formant frequencies (F_1 & F_2) for the final steady-states of

diphthong productions with "sustained utterances of /I-i/ and /U-u/" (p.592). Wise's results are noted in Table II although it is not clear what exactly Wise intends in the passage just cited.

The comparisons reported in Table I for the initial portion of the diphthong show that the studies were in agreement only for the diphthong /ɔI/. The initial portions of the diphthongs /aU, aI/ were comparable to several monophthongal vowels. In another study, not reported in the table, Gay (1968) noted that onset positions for diphthongs were preserved across changes in rate, but were "not clearly identifiable" (p.1571). Dolan and Mimori (1986) in a partial replication of the Gay study suggested that there is a tendency for the initial element of a diphthong to be centralized when diphthongs are produced at a fast tempo.

Overall, the comparisons reported in Table II indicate that there is variation in the monophthongal vowels with which final portions of diphthongs are closest in terms of formant values. Wise (1965) noted that the terminal frequencies of /oU/ were, on differing occasions, comparable to any one of three monophthongs (/U,u,o/). In contrast, Lehiste (1964) noted that terminal frequencies of /oU/ were not comparable with either /U/ or /u/. In addition, even though the diphthongs /aU, oU/ and /aI, eI, ɔI/ are sometimes transcribed as ending in the semi-vowels /w/ or /y/, both Lehiste (1964) and Wise (1965) noted that the terminal steady-states for these diphthongs were not comparable in formant frequencies to these semi-vowels. Gay (1968) reported that at a fast tempo final target vowels were not reached. The diphthongs /aI, eI,ɔI/ showed higher first formant and lower second formant

offsets as phoneme duration decreased while the diphthongs /aU,oU/ showed higher F_1 and F_2 offsets. Dolan and Mimori (1986) reported the same pattern as Gay for F_2 offsets (F_1 offsets were not mentioned).

From the results reviewed above, it appears that there is, in general, variability in respect to the vowel positions at which the various diphthongs begin and end. In particular, no one diphthong can be distinguished as consisting of a singular sequence of monophthongs. Some of this variation may be due to the use of speakers from different dialects, a point which cannot be established from the studies cited in the tables. Nevertheless, an optimal scheme for discriminating among these diphthongs should accommodate this variation. Further, even if initial and final portions of diphthongs are comparable to particular monophthongs in formant values, it is not clear whether it is legitimate to identify these portions with those monophthongs given their differences in other acoustic properties. Thus, Lehiste (1964) noted that the diphthongs /aU,aI,ɔI/ exhibited durations which were comparable to those of intrinsically long monophthongs and not to a summation of corresponding monophthongs (cf. pp.185-186 of that work for other considerations). Such differences may explain the comment in Lehiste and Peterson (1961) that "neither of the elements comprising the diphthongs is ordinarily phonetically identifiable with any stressed English monophthong" (p.276). Ladefoged (1982), discussing the auditory quality of initial and final portions of

diphthongs, makes a perhaps similar remark: "... contrary to the traditional transcriptions, the diphthongs often do not begin and end with any of the sounds that occur in simple vowels" (p.76).

1.3 F_2 rate of transition

Another basis for acoustically distinguishing diphthongs is suggested by the results of Gay (1968). In this study, the glide component of the diphthongs /aU,aI,eI,oU,ɔI/ was identified with (or: coincident with) the course of movement in the second formant. Particular diphthongs were found to show little variation in the rate of F_2 transition (in Hz/ms) across changes in tempo. Moreover, the mean values reported for the F_2 rate of transition appear to fall into distinct ranges for each diphthong. This suggests that these diphthongs may be distinguishable, at least in part, with respect to this parameter. Dolan and Mimori (1986) however found that F_2 rate of transition was in general not invariant across changes in tempo; the F_2 rate of transition was found to increase significantly with increased tempo for the diphthongs /aU, eI, oU, ɔI/.

1.4 Overall spectral properties of diphthongs

In Holbrook and Fairbanks (1962) the overall formant pattern displayed in the course of a particular diphthong production was considered without segmentation into steady-state and glide components. Instead, frequency and amplitude measurements for

the first three formants of a given production were made from five sampling points along a spectrogram. Median frequencies of F_1 and F_2 at each of the successive sampling points were obtained from spectrographic displays of a given diphthong and then plotted as points in a F_1 by F_2 space (Koenig scale). The successive points representing a given diphthong were then connected, forming a path within the space. Examination of these paths permitted Holbrook and Fairbanks to establish three classes of diphthongs. One class of "diverging diphthongs", comprising /aI,eI,ɔI/, was characterized by a falling first and a rising second formant. A class of "parallel diphthongs", consisting solely of /oU/, displayed a decrease in both formants with F_1 and F_2 maintaining a constant ratio. The phonemes /aU/ and /ju/ were considered "converging diphthongs". That is, the values for the first two formants were found to converge in the course of the productions of these phonemes. For the phoneme /aU/ the values of F_1 and F_2 both fell, the second more sharply than the first. For the phoneme /ju/ the value of the first formant rose slightly whereas the second formant descended considerably. It may be mentioned here that /ju/ is often distinguished from the other phonemes as an "ongliding diphthong". That is, it is usually considered to have a nucleus steady-state which follows rather than precedes the glide segment. Ladefoged, in his A Course in Phonetics considers /ju/ as a diphthong on the basis of historical considerations (its development from a vowel) and simplicity of phonological statement.

1.5 Theoretical background: The present study

The present investigation follows Holbrook and Fairbanks (1962) in examining the course of spectral change for productions of the phonemes /aU,aI,eI,oU,ɔI,ju/. These phonemes were considered in two contexts and in two rate (slow, fast) and two stress (stressed, unstressed) conditions. After a number of preliminary efforts, it was decided not to attempt segmentation of the productions into steady-state and glide components (cf. Gottfried 1989). This decision was based upon the methodological difficulty of establishing objective guidelines for delimiting such segments, and the theoretically unresolved issue of the perceptual significance of segments so identified (Bladon, 1985; Bond, 1978, 1982). These factors led us to evaluate methods for characterizing the diphthongs in terms of their overall spectral patterns, instead of steady-states and glides, within the auditory-perceptual space of Miller (1989). The auditory-perceptual space provides a tool for examining the formant trajectories of diphthong productions and establishing their distinctive acoustic properties. Within the space, differences presented by talker, age, and gender are normalized by consideration of formant ratios and their relation to a low-frequency reference.

In the auditory perceptual space, the values of the first three significant spectral prominences (denoted as SF1-SF3) together with the fundamental frequency for each millisecond of a

given production are transformed into values of three coordinates defined by the following equations:

$$x = \log(SF3/SF2),$$
$$y = \log(SF1/SR),$$
$$z = \log(SF2/SF1)$$

where SR is a reference frequency, given by the following equation:

$$SR = 168(GMF0/168)^{1/3}$$

and GMF0 is the geometric mean of the speaker's F_0 for the production.

In Miller (1989) it is noted that the monophthongal vowels of American English fall into a narrow slab, called the "vowel slab". The axes of the space can, by the following equations, be altered so that this slab is brought into a vertical position:

$$x' = .7071\ (y-x)$$
$$y' = .8162(z) - .4081(x+y)$$
$$z' = .5772\ (x+y+z)$$

Diphthongs, as vocalic speech sounds, would also be expected to fall within this slab. Furthermore, when retroflexion or nasalization do not occur, most of the variation of vocalic

spectra are captured within the x′ and y′ axes. Under these circumstances only small variations in z′ are observed, which are related to lip rounding. At the risk of losing some indications of lip-rounding or of changes in lip-rounding in the diphthongs /aʊ̈,oU,ju/, the analyses and graphs presented here use the x′y′ axes.

The values of F_0 and the first three spectral prominences for each ms yields a point within the space; successive points over the course of a production yields what is termed a "sensory path". The distance between two given points along a path within the space indicates the amount of spectral change between the two formant patterns represented by these points. For purposes of illustration, the sensory path for /ɔI/ produced as an isolated citation form is displayed in Figure 1. In the figure, triangles, oriented in the direction of the path, are positioned along the path every 50 ms indicating the extent of spectral change over successive intervals. In the research presented below we obtained sensory paths such as the one illustrated for the diphthong tokens. It was hypothesized that intended diphthong productions could be classified by selecting appropriate features of these paths as parameters. Moreover, Miller (1987, 1989) and Miller and Chang (1989) forwarded the hypothesis that perception of diphthongs is signalled by their paths within this space. Identifying appropriate parameters for the classification of intended diphthong productions would assist in more precisely formulating hypotheses as to the perception of diphthongs.

[insert figure 1]

2. Methods

2.1 Speakers and Speech Material

The subjects were four native speakers of Midwestern American English, two male and two female, with no known speech or hearing disorders. Two lists of words were constructed containing the six diphthongs in two contexts: [b_d] and [h_d]. To minimize the number of non-English words used, the first context for the phoneme /ju/ was [b_t], an exception yielding the slang form "beaut". A carrier sentence was constructed in which each of the target words is preceded by the word "kay". In one listing of the carrier sentences, this word was displayed in upper-case and subjects were instructed to produce the sentence with stress on that syllable. In another listing of the carrier sentences, the word "kay" was exhibited in lower-case: subjects were in this case instructed to pronounce the sentence with stress on the target word. Lists of the sentences were obtained by randomizing six repetitions of each of the words for each of stress and rate conditions and for each speaker. Table III provides the format of the carrier sentence and the word lists. In one of the recording sessions subjects were instructed to produce the sentences in a deliberative, careful manner (slow); in the other, the instructions were to speak as rapidly as comfortable without

making mistakes (fast). Only four of the six repetitions were used in the analysis. If a particular utterance was heard as containing extraneous noise, e.g., shuffling of papers, or presented problems in the determination of formant values (see below), another of the available repetitions was employed. This procedure yielded 768 utterances (2 (rate) x 2 (stress) x 2 (contexts) x 6 (diphthongs) x 4 (repetitions) x 4 (subjects)).

2.2 Recording and Digitization

Each speaker read the lists in an anechoic chamber with the microphone placed at a height equal to and 1/2 meter away from the mouth, using a low-noise microphone/preamplifier combination (Bruel and Kjaer 4179/2660) . The speakers were instructed to speak in a normal conversational manner which resulted in a signal of 70 dBA at the microphone. Output from the microphone was channeled directly into a digital audio recorder (Sony PCM-501ES) operated in its 16 bit mode. The frequency response of the total system was within +/- 2 dB over the range from 10 Hz to 10 kHz. The signal-to-noise ratio was >65 dB A-weighted.

Immediately prior to the recording, subjects were requested to begin a preliminary recitation of the utterances, so that the experimenter could set the recording level to about -9 dB VU. After setting this level, the experimenter recorded a calibration tone. The calibration tone, generated by a synthesizer (Hewlett-Packard 3325A), consisted of a 1-kHz sine wave and was kept at a constant output level of 69.5 mv (equal to 70 dB SPL at the

microphone 1V/Pa). The recording of the calibration tone allows calculation of the SPL of the subject's speech.

Digitizations of four of the six repetitions of each test word were made at 20 kHz with 16-bit precision, through a Digisound-16, a stimulus access processor and both a 50 Hz analog high-pass filter (to remove incidental noise) and a 10 kHz anti-aliasing filter. Any residual AC noise or "hum" was removed by digitally notch filtering the files at 60 Hz after digitization. The files were stored on a MicroVax II or VAXstation 3200 computer to facilitate further processing by the Interactive-Laboratory-System (ILS) commercial software package.

2.3 Measurements - Durational

Three types of durational measures were made from the speech wave as displayed on a graphics terminal (Hewlett-Packard, Model 2623). These measures were obtained using a hardware/software system, which allows cursors to be placed at significant points along the waveform and measurement of the marked durations to the nearest millisecond. The three durational measures were as follows. (1) Total sentence duration was measured from the onset of the first glottal pulse marking the onset of voicing for the word "I" to the end of the nasal murmur in the final word "again". (2) Target-word duration was measured (a) in the context [b_d] from the onset of the initial closure following the syllable "kay" until the end of the final closure, and (b) in the context [h_d] from the onset of [h] frication until the end of

the final closure. The onset of the [h] frication was sometimes difficult to locate as the [h]s were generally voiced. Most often the onset was taken as the point at which significant frication was first observed in the waveform. In other cases, it was necessary to listen to various windowed sections of the waveform to assist in determining the onset of the [h] sound. (3) Vocalic-nucleus duration was measured from the end of the initial consonant of the target word ([b] or [h]) until the onset of closure for the final consonant ([d] or [t]).

2.4 Measurements - Spectral

Following digitization, the waveforms were modified so that all but the isolated vocalic nucleus was set to zero. Analysis was made using the API (analysis with pitch extraction) command of ILS, with a 24 ms Hamming window moving in 1 ms steps, 24 poles, and a pre-emphasis factor of 98%. This command performs a linear-prediction coding (LPC) on the digitized waveform and a cepstral analysis for the determination of the fundamental frequency. A fast Fourier transform (FFT) is performed on the LPC autoregression coefficients by means of the spectrogram display (SGM) command of ILS in order to obtain values for the first three formants. The values for F_0-F_3 were stored in a tabular-format file which could be hand-edited. In cases where the cepstral analysis did not yield values for the fundamental frequency, these values were obtained by determining the time interval for three pulse periods, and then calculating the F_0

values as the inverse of the period determined. In cases where any of the first three formants were missing for brief intervals, direct FFT's were made on the waveform in order to determine appropriate formant values to be inserted into the file. In cases where the LPC analysis did not yield values for formants over large intervals, the root-solving command of ILS was used to obtain formant values for the token. This situation arose most frequently for those productions during which F_2 and F_3 were close in value (i.e., for tokens of /aI, eI, ɔI, ju/). Formant values for more than half the tokens of /eI/ and /ju/ were obtained by means of this root solving command.

2.4.1. Determination of Initial and Final Transitions.

In the preceding discussion, the term "vocalic nucleus" has been employed to refer to the diphthongal segment inclusive of transitions from and to the environing consonants. Transitions from the preceding [b] or [h] to the diphthongal element are here termed "initial transitions"; transitions into the final consonant are "final transitions". The end of the initial transition delimits the diphthong "onset" whereas the starting point of the final transition delimits the "offset" of the diphthong. Determination of the onset and offset of each intended diphthong production was made from a graphic display of the formant patterns on a log-frequency scale. Figure 2 (top panel) exhibits a sample formant pattern in which the onset and offset of a diphthong production is identified. The formant pattern shown is for a production of a stressed /aU/ in the

context [h_d]. The onset is taken as the point at which the ratio between F_1 and F_2 becomes constant; in this case, this was the first point in the vocalic segment as indicated by the leftmost vertical line. The offset is taken here as the latest point within the last 30 ms of the segment which just precedes a noticeable increase in the value of F_2, shown in the figure by the rightmost vertical line. In establishing the onset and offset for tokens of the diphthongs, these times were estimated to the nearest multiple of 5 ms with respect to the onset of the vocalic segment. The manner in which the onset and offset values of diphthongs are identified depends on the consonantal context and other features of the particular diphthong in question. The guidelines which were employed for the determination of these points are provided in Appendix A.

2.4.2 Generation of Sensory Paths.

While identification of the onset and offset of the diphthong was determined from an examination of the formant plots, it was hypothesized that an appropriate characterization of the diphthongs could be obtained from their associated paths as displayed in the auditory-perceptual space. These sensory paths were generated as follows. Each formant file was smoothed with a first-order resonator with a center frequency of 20 Hz as this serves to eliminate minor perturbations in the formant trajectories. Then, the values of F_0 and the first three formants for each successive millisecond were converted into

points within the auditory-perceptual space by means of the equations presented above (cf. section 1.5). The resulting sensory path provides an integrated display of the course of spectral change. Figure 2 (Middle panel) shows the smoothed version of the formant pattern corresponding to the top panel of figure 2. The sensory path generated from this smoothed formant plot by the equations noted above is shown in the bottom panel of figure 2.

[insert panels of figure 2 about here]

As noted in section 1.5, it was hypothesized that various features of sensory paths obtained for the diphthong tokens would serve as a basis for classifying the intended production.

3. Results

3.1 Durational Results

In this subsection, the results of the durational measures are reported along with analyses of variance performed using subjects as replicates in a 4 factorial design with 48 cells: 2 tempos x 2 stresses x 2 contexts x 6 diphthongs. The F-ratios and degrees of freedom are reported only for significant effects ($p<.05$). The analyses indicated a number of interactions between the identity of the intended diphthong and tempo, stress, or context. It was noticed that many of these interactions involved the diphthong

/ju/, which, in the context of a preceding [b] was followed by [t], yielding the exceptional form "beaut". In those cases where these interactions were due solely to the inclusion of /ju/ in the data set (as confirmed by separate analyses of variance excluding these tokens), the interaction is not reported.

3.1.1 Sentence durations.

In Table IV the mean duration (in ms) for sentences in the different tempo-stress conditions is displayed. Sentences spoken at a fast tempo were shorter in duration than those spoken at a slow tempo by 25%. The effect was significant [$F(1,3)=17.84$; $p<.03$]. This result indicates that subjects did change their rate of speech as instructed. There was also a three-way interaction among stress, context, and identity of the diphthong [$F(5,15)=3.86$; $p<.02$].

3.1.2. Target syllable durations.

In Table V the mean durations (in ms) for the target syllables pooled across diphthongs are provided. Target syllables produced at a fast tempo were 23% shorter than those produced at a slow tempo. While sizeable, the effect only approaches significance [$F(1,3)=7.75$; $p<.07$]. Target syllables in the unstressed condition were 27% shorter than those in the stressed condition; this effect was significant [$F(1,3)=24.35$; $p<.02$]. The duration of the target syllable was 8% shorter in the [h_d] context than in the [b_d] context. While the difference was small, the effect of context was highly significant [$F(1,3)=60.65$; $p<.01$].

3.1.3. Vocalic segment durations.

Tables VI - VII provide mean durations (in ms) for the vocalic segments containing the various diphthongs, i.e., the diphthong inclusive of transitions from and to the environing consonants. Vocalic segments produced at a fast tempo were 24% shorter than those produced at a slow tempo. While the effect was not significant, it should be noted that its size, about 25%, is consistent with that displayed by sentence and word durations. A nearly equivalent reduction of duration was seen in the effect of stress: mean durations of the various vocalic segments when unstressed were 26% shorter than when stressed. This effect was significant [$F(1,3)=17.05$; $p<.03$]. Vocalic segments in the [h_d] context were about 12% shorter than those in the [b_d] context; this effect of context was significant [$F(1,3)=14.19$; $p<.04$]. Vocalic segments requiring less articulatory movement (/eI, oU, ju/) were 9% briefer than those which require greater articulatory movement (/aU, aI, ɔI/). This effect of the identity of the diphthong was statistically significant [$F(5,15)=38.00$; $p<.01$]. This pattern was preserved in each rate condition and is in agreement with Holbrook and Fairbanks (1962) and Gay (1968).

There were three interactions of note: between tempo and identity of the diphthong; between context and identity of the diphthong; and between stress and context. The difference in duration between vocalic segments produced at a slow tempo and those produced at a fast tempo was highest for /aI/ and /aU/ (27%) and least for /ju/ (19%). The differences in duration

between the slow and the fast tempo for the other diphthongs are intermediate (approximately 23%). The interaction between tempo and identity of diphthong described above was significant [$F(5,15)=6.5$; $p<.01$]. A similar pattern among the diphthongs was found in respect to context. While all vocalic segments were shorter in duration in the [h_d] context than in the corresponding [b_d] context, the differences were highest (17%) for /aU, aI/ and were about the same (12%) for the other vocalic segments. In contrast, the duration of /ju/ was shorter in the context [b_t] than in the [h_d] context. The interaction between context and identity of diphthong was highly significant [$F(5,15)=18.00$; $p<.001$]. Unstressed vocalic segments were 31% shorter than stressed segments in the [b_d] context whereas in the [h_d] context unstressed segments were 24% shorter in duration. This interaction between stress and context effects was statistically significant [$F(1,3)=12.41$; $p<.04$].

3.1.4. Summary.

In summary, subjects did alter their speaking rate as instructed. While there were differences among subjects, the effect of tempo was consistent for each subject at the level of sentence, target syllable and vocalic segment. For three of the subjects sentence durations in the fast tempo were reduced by approximately 15% relative to the slow tempo, while for the fourth, this reduction was 37%. For each subject target-syllable and vocalic segment durations were reduced by approximately the same percentages. Subjects also placed stress on the target- or dummy-syllable as

instructed, as can be noted from the signficant effect of stress on the duration of target words and vocalic segments.

3.2. Parameters for Classification

Previous consideration of the diphthongs of American English as sensory paths within the auditory-perceptual space led Miller and Chang (1989:34-43) to hypothesize that diphthongs could be characterized in respect to the regions in the space in which they originate and the direction of their movement. In the research reported here we investigated this hypothesis by establishing various parameters in the space as a basis for classifying the intended diphthong productions. The spectral patterns at the onset and offset of the diphthong were used as "anchor points" in identifying these parameters. An illustration of the onset and offset points along the sensory path of a particular diphthong production is provided in the bottom panel of Figure 2, where the onset and offset are marked by crosses. The following four parameters were identified in respect to two such points in the space (themselves identified as two pairs of x'and y' coordinate values). Two of these parameters were the x' and y' values for the onset (hereafter, x',y'). The other parameters were established in respect to both points, onset and offset. The "distance" (d) between the two points was calculated as the length in log units of the straight-line connecting them. This line is referred to below as a "diphthong line". The angle (θ) for a particular token is established as that formed between

the vertical axis erected, parallel to the y′ axis, at the x′ value at onset and the diphthong line proceeding clockwise (see figure 2, bottom panel for an example). As will be illustrated later, the angle provides a measure of the pattern of formant movement within the space.

In the next three subsections, each of the parameters is examined individually. We then consider the success of these parameters as a means of classifying the intended diphthong productions.

3.2.1. Values of onset (x′, y′).

In figure 3 the mean x′ and y′ coordinate values are provided for the onset of each diphthong across all conditions of context, rate, and stress. For each diphthong the vertical bars indicate one standard deviation above and below the mean y′ and the horizontal bars indicate one standard deviation to either side of the mean x′ value. Both /eI / and /ju/ have onsets that are well separated from each other and from onsets of the other diphthongs. On the other hand, the pairs /aU, aI / and /oU , ɔI/ are not well separated. The onset values for /aU/ are larger for both the x′ and y′ axes than the onset values for /aI/. Similarly, the onset values for /oU/ are larger for the x′ and y′ axes than those of /ɔI/. The range of x′ values for /aI/ indicated by the horizontal bars is wider than that for /aU/ and nearly covers the latter. For /oU/ and /ɔI/, there is only partial overlap among these values. The overlap of y′ values for

the pairs /aU,aI/ and /oU,ɔI/ extends roughly one standard deviation, coinciding with the lower range for /aU/ and /oU/.

It is also of interest to consider onset positions in terms of the effect of rate and stress. Dolan and Mimori (1986: 142-44) suggest that "the first element of a complex vowel tends to move toward the center of the vowel space as rate is increased" (for contrary results, see Gay 1968). This claim was evaluated in the following manner. First, we established a point in the auditory-perceptual space which corresponds to that of the neutral reference vowel described in Fant (1970). This neutral reference vowel is associated with a uniform vocal tract, 17.6 cm long, in which the first three resonances are 500, 1500, and 2500 Hz respectively and the fundamental is 133 Hz (for a male speaker). [Note: The point within the space which corresponds to the neutral reference vowel produced by a female uniform vocal tract with a typical female F_0 of 230 Hz is identical to that of the male]. We then calculated the (Euclidean) distance between the x' and y' values for the onset of each diphthong token and this "neutral point". The mean (Euclidean) distance from this point of each diphthong onset in each rate-stress condition is shown in Table VIII. The results of pair-wise t-tests on the means, pairing each rate-stress condition with all of the others, is displayed in Table IX. Significant results were obtained for pairs differing in stress condition as well as for pairs differing in both stress and rate: SS/FU and FS/SU. This suggests a tendency toward centralization, though the effect is promoted more by absence of stress than by an increased rate.

3.2.2. Distance (d).

In Table X the mean distances (log units) traversed by each diphthong under the various rate and stress conditions are displayed. This measure provides an indication of the overall spectral change associated with each diphthong (cf. section 1.5). With the exception of /oU/, each diphthong courses progressively less distance as one proceeds from the slow-stressed condition to the fast-unstressed condition. As might be expected from their articulation, the diphthongs /ɔI/ and /aI/ traverse the greatest distance, /eI,oU/ traverse the least, and /aU,ju/ occupy an intermediate position. Pair-wise t-tests on the means, pairing each rate-stress condition with each of the others, shows significant differences for all but the comparison of the two unstressed conditions (Table XI). That is, there is a significant reduction of distance travelled for unstressed diphthongs in comparison with their stressed counterparts as well as when the tempo alters from slow to fast for stressed diphthongs. The latter result is consistent with those of Gay (1968) which indicated that diphthongs produced at a fast rate do not reach their "target" (i.e., travel less distance in an F_1 by F_2 space than those produced at slower rates).

3.2.3 Angle (Θ).

Table XII provides the means and standard deviations of the angle parameter for each diphthong within each of the rate-stress groups. For each of the diphthongs, there appears to be little change in the mean angle across the rate-stress groups (paired t-

tests revealed no significant differences). As discussed above, angles are formed by the y' axis and the diphthong line, the straight-line between the points representing the diphthong's onset and offset. On the average, these lines for a particular diphthong are similarly oriented in the auditory-perceptual space regardless of stress and rate conditions (cf. section 4). Examination of the standard deviations for each diphthong both within and across the various rate-stress conditions reveals two patterns of note. Firstly, within each of the conditions there are generally progressively larger standard deviations for the angle parameter following the order /ɔI,aI,ju,eI,aU,oU/. Secondly, except for /ɔI,ju/, there appears to be a tendency for standard deviations of individual diphthongs to increase across the rate-stress groups displayed in the table. Figure 4 provides another display of the variability exhibited by particular diphthongs with respect to the angle parameter. For each of the diphthongs the range of angular values falling within the 25th to the 75th percentile is indicated on a pie-chart. The diphthongs /aU,ɔI,ju/ are seen to have unique ranges. The range for /aU/ exhibits a westerly orientation in the third and fourth quadrants. The range for /ɔI/ is oriented in a northeasterly direction; the range for /ju/ is oriented in the opposite direction. The remaining diphthongs, /aI,eI,oU/, have ranges which overlap wholly or in part. All of them share a roughly northern orientation. In the auditory-perceptual theory, movement of a path in this northerly direction indicates divergence of the first two formants. For the diphthongs /aI,eI/

this pattern of formant movement agrees with earlier observations, e.g., by Holbrook and Fairbanks (1962), whereas for /oU/, it does not. Three hypotheses may be offered for the exceptional status of the diphthong /oU/. Firstly, by considering only x′ and y′ values of diphthong onsets and offsets in the determination of angle, we have ignored values with respect to the z′ axis. In Miller (1989) it is hypothesized that the effect of rounding is indicated by movement within the vowel slab from larger values of z′ to smaller values. If productions of /oU/ are characterized by an increase in rounding from beginning to end, then a more accurate indication of angular movement for tokens of /oU/ would have to consider z′ values as opposed to a simple projection of the paths onto the x′and y′ axes. That is, it is possible that the simple projection inaccurately indicates a significant northerly movement in the space whereas the significant direction may be back along the z′ axis. Secondly, some of the tokens of /oU/ may have been produced as monophthongs in which case the angle of movement may be irrelevant, and the data were not screened for this possibility. Finally, the determination of final transitions (and, hence, of the diphthong offset) in the case of tokens of /oU/ may be in error. As noted above, the offset was chosen as the latest point within the last 30 milliseconds of the segment from which values for F_2 noticeably increase or values for F_1 decrease. In examining formant plots of /oU/ tokens, it was frequently observed that values for the second formant increased prior to the last 30 milliseconds of the vocalic segment. If the point

from which F_2 first noticeably increases were chosen as the diphthong offset, one might obtain angular values which are in accord with results of other investigations, i.e., values indicating a generally westerly movement similar to that of /aU/ in figure 4. All three of these hypotheses deserve further investigation.

3.3 Classification results

3.3.1 Major Results.

While other investigators have noted various properties which might distinguish particular diphthongs (cf. section 1), these properties have not been evaluated as parameters within a classification procedure. One of the principal aims of this study is to evaluate the success of various parameters in classifying productions of American English diphthongs. The parameters, as noted above, were the x′ and y′ coordinates of the diphthong onset (x′, y′), the distance (d) which roughly indicates the extent of formant movement, and the angle (θ), indicating the pattern of formant movement. Linear discriminant analyses (LDA) were employed in order to assess how well particular combinations of these parameters served as classifiers of the productions. This procedure calculates the group mean for each decision variable and then assigns each token to a group on the basis of an a posteriori probability of group membership. Input to the procedure consisted of values of one or more of the four parameters identified above.

There is a problem in accomodating angular measures within this procedure. For those diphthongs whose angular range includes tokens within 270° to 360° as well as tokens within 0° to 90°, simple calculations of group means would provide an inaccurate depiction of variations in angle. Accordingly, angular measures of tokens of /aI, eI, oU/ within the range of 0° to 90° were increased by 360 degrees before submitting these values to analysis.

In the initial pass, the tokens were assembled into four groups corresponding to each rate-stress condition. Each group comprises 192 tokens. Using the resubstitution (R) method of classification, linear discriminant analyses were run for each possible combination of parameters. The results for those combinations yielding percentages greater than 80% for all four groups as well as for the total data set are reported in Table XIII.

From this table it can be seen that performance for the various parameter combinations is higher for the stressed groups than for the unstressed groups and is highest for the slow-stressed group. The poorest performance is for those tokens in the fast-unstressed group. In the unstressed condition, as noted above, distance is reduced and diphthong onsets tend to be more centralized. Consequently, we would expect that these parameters would be less effective in correctly classifying tokens in this group.

To further appraise the success of these parameters, the classification matrix obtained for one group of the tokens, i.e.,

the slow-stressed group, was used to classify those in the other groupings (referred to below as the "second pass"). Table XIV reports the percentage success and a posteriori probabilities for those parameter-combinations noted in the previous table. As in the initial pass, correct classification for the stressed groups is generally higher than correct classification in the other groups and performance for the fast-unstressed group is consistently poorest. On the average, the two best performing combinations are: x′, y′, d, θ and x′, y′, θ (discussed further in section 4).

3.3.2 Other Parameters.

The results presented in Tables XIII-XIV indicate the overall success of various parameter combinations in classifying diphthong productions. Results gleaned from other combinations, not presented in these tables, are nonetheless of interest, as they illustrate the manner by which particular parameters contribute to the overall success. Below we briefly consider the following parameters: (a) the diphthong onset (x′, y′) , (b) distance, and (c) angle (θ) in both the initial pass (with the R-method of classification) and the second pass (with the classification matrix derived from the slow-stressed group employed to classify the other rate-stress groups).

(a) Confusion matrices derived from specification of the diphthong onset exhibited the expected confusions between /aI/ and /aU/ and between /ɔI/ and /oU/. Nonetheless, this combination overall yielded an average 82% correct classification in the

initial pass and 77% in the second. These percentages are higher than might have been expected. The success of this combination shows that the onset of the diphthongs occupy specific though overlapping regions in the auditory-perceptual space.

(b) Confusion matrices derived from analyses in which distance served as the sole parameter were also in line with expectations. For instance, it can be seen from Table X that the mean straight-line distance covered by /aI/ and /ɔI/ was greater than that for the other diphthongs. This corresponds to the greater articulatory movement with which these diphthongs are produced. Using distance as a classifier, one would expect confusions between /aI/ and /ɔI/, as well as between /eI/ and /oU/; this is indeed what appeared in all conditions. Overall, distance as a parameter yielded an average 38% correct classification in the initial pass and 28% in the second.

(c) Confusion matrices derived from analyses in which angle served as the sole parameter were in line with expectations: tokens of /ɔI/ and /ju/ were generally correctly classified and there were frequent confusions among /aI/ and /eI/ (cf. Figure 4). Angle served as the best single classifier in both passes, yielding an average 60% correct classification (across the rate-stress groups) in the initial pass and an average 58 % correct classification in the second. These scores are higher than those with x′ and y′ taken singly, though, as noted above, the combination of x′ and y′ performed quite well.

Examination of confusion matrixes for the parameter combinations just considered illustrates the way in which

specific parameters contribute to the collective success of the parameter combinations recorded in Tables XIII- XIV. For instance, addition of either distance or angle to the x',y' parameter combination discriminates between /aI/ and /aU/ and between /ɔI/ and /oU/.

3.3.3 Summary.

In summary, using only values for the onset and offset of each diphthong token within the auditory-perceptual space, a set of parameters can be established which effectively classifies the intended production under the conditions of this study. Establishing these parameters does not require division of the vocalic segment into steady-state and glide portions. While particular combinations of parameters are seen to be more successful in obtaining correct classifications of individual diphthongs, it is, in the authors' opinion, the overall excellent performance of these combinations which is of most significance. The import of particular combinations and their possible bearing upon perception of diphthongs is addressed in the latter part of the next section.

4. Discussion

The results presented above are in accord with the general observations made and various groupings proposed for these diphthongs in other studies. In respect to duration, the phonemes

/eI,oU,ju/ were generally shorter than /aU, aI,ɔI/. The phonemes /eI, oU, ju/ had an average duration of 147 ms (s.e.= 2.30, N= 384)) across all conditions of rate and stress, while /aU,aI, ɔI/ had an average duration of 171 ms (s.e.= 2.80, N= 384) across these conditions. This is in general agreement with Holbrook and Fairbanks (1962) and Gay (1968). The diphthongs may also be assembled into various groups on the basis of the typical direction of F_1 and F_2 movements over the course of the production. Comparisons of these movements (in a F_1 by F_2 space) with those in the auditory-perceptual space are somewhat complicated, given the dependence of the x′ and y′ coordinates upon the third formant and the sensory reference (cf. section 1.5). As noted in the Introduction, the /-I/ diphthongs were characterized by an increasing second formant and a falling first formant, termed "divergent" in Holbrooks and Fairbanks (1962). Figures 5 - 7 display formant trajectories for sample /-I/ diphthongs (left panels) paired with their associated sensory paths (right paths). As can be seen, divergence of the first two formants corresponds to upward movement in the space. The /-U/ diphthongs were characterized by decreasing first and second formants. The diphthong /aU/ displayed in the panels of Figure 2 and /oU/ displayed in those of Figure 8 exhibit leftward movement in the space corresponding to the decreasing values in the two formants. The phoneme /ju/ in Figure 9a shows an increase in the first formant and a decrease in the second, which corresponds to the downward movement seen in its associated path (Figure 9b).

[Insert Figures 5,6,7,8, & 9]

Summarizing the discussion in section 3.2 it was noted that whereas distance coursed by each diphthong is reduced as tempo is increased and the onset of each diphthong tends to be centralized when not stressed, the pattern of formant movement, as distinguished by angle, tends to remain constant across the various rate-stress conditions. This suggests that diphthongs are discriminable in respect to the region of onset together with the direction of formant movement (while the extent of movement may vary). This supposition is borne out by the results of the linear discriminant analyses (section 3.3). In the analysis in which the classification matrix for the slow-stressed tokens was used to classify tokens in the other rate-stress groups (Table XIV) angle and onset appear in the the two best performing combinations of parameters: (a) x',y',d, θ and (b) x', y', θ . The first combination yields an overall average of 93% correct classification; the second, 91% correct classification. Addition of distance as a classifier boosts performance only slightly. Accordingly, the following discussion will focus on the latter combination.

As noted in section 3.3.2, angle served as the best single classifier in both passes (i.e., using the R-substitution method in the first and the classification matrix derived from the slow-stressed group in the second), with an average 60% correct classification for the first pass and an average 58% in the second. Specification of the diphthong onset yielded an average

82% correct classification on the first pass and 77% on the second. Examination of confusion matrices for both runs of linear discriminant analyses reveals that those diphthongs confused in respect to angle are discriminated in respect to onset and vice versa.

These results suggest that "onset regions" together with the direction of formant movement are the perceptually salient factors in identifying and discriminating these phonemes. "Onset regions" for the diphthongs, identifiable in respect to clusters of points in the auditory-perceptual space, need not correspond to any monophthongal vowel of English (cf. section 1.2). The phonemes /eI/ and /ju/ appear to have relatively distinctive "onset regions", whereas those for the pairs /aU,aI/ and /oU,ɔI/ appear to have greater overlap (figure 3, section 3.2).

In terms of the three-dimensional auditory-perceptual space, an "onset region" extended in the the direction of formant movement within the space delimits a "pipe". Miller (1987) and Miller and Chang (1989) suggest that the perception of glides and diphthongs may be induced by movement along a path through a "pipe" with various conditions attached to such variables as, e.g., point of entry and exit, extent of movement, and velocity of movement (see Chang 1987 for an investigation along these lines). The present study yields support for this hypothesis and provides preliminary indications as to the relevant features and conditions associated with "pipes" for diphthongs, e.g., that extent of movement is dependent upon speaking-rate.

In future work we intend to address the issue of more precisely delimiting diphthong onsets and offsets and to evaluate the perceptual import of the parameters discussed above. The results reported for linear discriminant analyses should first be compared with perceptual evaluations of the productions. The parameters might also be evaluated in respect to alternative "offset" points stationed along a sensory path for various intervals of time and/or distance. Comparison of linear discriminant analyses performed in this manner with corresponding perceptual tasks would provide useful information on the requisite kinds and amounts of formant movement which induce perception of diphthongs.

Acknowledgements

The authors are indebted to Caroline Monahan for assistance on statistics and Aaron Schlafly for efforts in data collection. Marios Fourakis and John Hawks provided useful discussion throughout the course of the research, and, together with Ira Hirsh, instructive comments on earlier versions of this paper. This research was supported by NIDCD Grant R01-DC00296 to the Central Institute for the Deaf.

References

Bladon, A. (1985) Diphthongs: A case study of dynamic auditory processing, Speech Communication, 4, 145-154.

Bond, Z.S. (1978) The effects of varying glide durations on diphthong identification, Language and Speech, 21, 253-263.

Bond, Z.S. (1982) Experiments with synthetic diphthongs, Journal of Phonetics, 10, 259-264.

Chang, H. (1987) SWIS: See what I say, a speaker-independent word recognition system by phoneme-oriented mapping on a phonetically encoded auditory-perceptual speech map. Washington University dissertation.

Dolan, W. and Mimori Y. (1986) Rate-dependent variability in English and Japanese complex vowel F2 transitions, UCLA Working Papers in Phonetics 63, 125-153.

Fant, G. (1970) Acoustic theory of speech production. The Hague: Mouton.

Gay, T. (1968) Effects of speaking rate on diphthong formant movements, Journal of the Acoustical Society of America 44, 1550-1573.

Gottfried, M. (1989) Some acoustic properties of diphthongs, Journal of the Acoustical Society of America, 86 (Suppl.1), S123.

Holbrook, A. and Fairbanks, G. (1962) Diphthong formants and their movements, Journal of Speech and Hearing Research, 5, 38-58.

Ladefoged, P. (1982) A course in phonetics, New York: Harcourt Brace Jovanovich.

Lehiste, I. (1964). Acoustical characteristics of selected English consonants, Bloomington: Indiana University.

Lehiste, I. and Peterson, G. (1961) Transitions, glides, and diphthongs, Journal of the Acoustical Society of America, 33, 268-277.

Miller, J.D. (1987) Auditory-perceptual processing of speech waveforms. In Auditory processing of complex sounds (W.A. Yost and C.S. Watson, editors), pp.257-266. Hillsdale, NJ: Laurence Erlbaum Associates.

Miller, J.D. (1989) Auditory-perceptual interpretation of the vowel, Journal of the Acoustical Society of America, 85, 2114-2134.

Miller, J.D. and Chang, H. (1989) Patent #4820059, U.S. Patent Office. (49 figures, 58 columns).

Peeters, W.J.M. (1987) Acoustic structure and perceptual relevance of 'steady states' and 'glides' within formant trajectories of diphthongs, complex vowels, and vowel clusters. In European Conference on Speech Technology, volume 1 (J. Laver and M.A. Jack, editors), pp.42-46. Edinburgh, UK: CEP Consultants.

Peeters, W.J.M. (1989) Vowels, diphthongs, and vowel clusters: A quantitative dynamic approach through synthesis-by-rule. In Proceedings of the Workshop on New Methods in Dialectology (Amsterdam 1987).

Trager, G. and Smith, H.L. (1951) An outline of English structure, Norman: Battenburg Press.

Wise, C.M. (1965) Acoustic structure of English diphthongs and semi-vowels vis-a-vis their phonemic symbolization. In Proccedings of the fifth international congress of phonetic sciences (E. Zwirner and W. Bethge, editors), 589-593. Basel: S.Karger.

Appendix A. Determination of Initial and Final Transitions

The establishment of the initial and final transitions presents several difficulties. Firstly, in determining the initial transition into the diphthongal segment, we cannot presume that there is an initial steady-state: diphthong productions do not invariably display an initial steady-state (Gay (1968), Gottfried (1989), Peeters (1987, 1989)). Moreover, various researchers have presented different criteria for what constitutes a steady-state (Lehiste and Peterson 1961, Gay 1968). Finally, as noted in section 1.2, it is not the case that a steady-state if present can be associated with some one monophthongal vowel. In the absence of well-established criteria for determination of transitions, an effort was made to develop a procedure which did not presume that the diphthong had an isolable nucleus. Rather, we considered whether displays of the first two formants of the tokens exhibited movement in the direction of the "target" monophthong by which the intended diphthong is generally transcribed. The procedure described below also allowed us to revise our initial determinations of the initial and final transitions.

The initial and final transition associated with each token was determined from a graphical display of the formants with frequency indicated in log scale. The initial transition was taken to extend from the onset of the vocalic nucleus up until

the first point (estimated within the first 30 ms to the nearest 5 ms) at which either (i) F_1 reaches a maximum or (ii) the F_2/F_1 ratio first stabilizes. In cases where the second criterion is satisfied at onset of the vocalic nucleus, the onset of the diphthong is identified with the onset of the vocalic segment. These evaluations were checked by noting whether the putative initial transitions displayed F_1 and F_2 movements in the direction of the "target nucleus", i.e., the vowel by which the "nucleus" of the diphthong is usually transcribed. As noted above, it is not supposed here that a given diphthong has a nucleus identifiable as that particular vowel or that it has a "nucleus" element at all. For the diphthongs written as /aU/ and /aI/, this target is /a/; for the diphthongs /oU/ and /ɔI/, it is /o/; for /eI/, it is /ɛ/. For /ju/ the target position (not nucleus) is /j/. The expected movements of the first and second formants from initial [b] or [h], the latter both in the case where there is anticipation of the following vocalic element and where there is no anticipation, into each of these "nuclei" is shown in Table AI.

Transitions from the diphthong into [d] (or [t] in the case of [bjut]), i.e., final transitions, are taken to begin at the latest point within the last 30 ms of the vocalic segment from which F_2 rises and/or F_1 falls markedly. There are two exceptions to this guideline: (1) Many instances of vocalic segments containing /ju/ display increases in F_2 considerably before this point, and the onset of the transition is taken in these cases to begin with the first pronounced rise in F_2; and (2) If there is

neither a rise in F_2 nor a fall in F_1, the end of the diphthong is taken as the end of the vocalic segment.

Determination of initial and final transitions thus delimits the onset and offset of the diphthongal element (the former coinciding with the termination of the initial transition; the offset coinciding with the initiation of the final transition). There were a sizeable number of cases in which closely neighboring points satisfied one or another of the conditions presented above for the onset or offset of the diphthong. For instance, in establishing the onset of a diphthong production, the F_2/F_1 ratio may stabilize shortly after F_1 has reached its maximum. The procedure developed allowed for readjustment of particular onset and offset points if values for the angular parameter (cf. section 2.5.1) exhibited considerable variance from the expected value. These cases were established as follows. After establishing the angular values associated with each token, we then calculated for each speaker, context, stress-rate condition, the average of the four angular measures obtained and compared this with the average obtained globally. The global means and standard deviations (s.d.) associated with each of the six diphthongs were (in degrees): /aU/, X = 261.7, s.d. = 63.2; /aI/, X = 358.7, s.d.= 18.8; /eI/, X = 331.9, s.d.= 54.4; /oU/, X= 330.6, s.d.= 71.7; /ɔI/, X = 24.8, s.d.=11.7; /ju/, X= 194.0, s.d.= 27.1. If the range 0^o-90^o is established as quadrant I, and successive 90^o segments as quadrants II, III, and IV, it may be noted that for the diphthongs /aI, eI/ the mean +/- 1 s.d. covers portions of both the IV and I quadrants; in the case of

/oU/ it includes that of quadrant III as well. In calculating the the means for each speaker, context, and rate-condition for these diphthongs (/aI,eI,oU/) tokens whose values were in the Ist quadrant were increased by 360 degrees. If this average diverged from the global mean by greater than one standard deviation in either direction, the formant plots were reexamined and other values determined for the onset and offset of the diphthong following the criteria noted above. These reevaluations were often made for unstressed diphthongs /aU, aI/ and /ɔI/ produced at a fast tempo. Of the 768 tokens, onsets and offsets for 88 tokens (11% of the total data set) were reevaluated.

Figure A1 displays both the original and the revised diphthong onset of a production of an unstressed /aI/. The initial onset was taken at the point at which F_1 reached a maximum; the revised onset was taken at the point from which F_1 and F_2 maintain a constant ratio.

Table I. Closest monophthongal vowel in comparison with initial portion of diphthong

Study	Diphthong					
	aU	aI	eI	oU	ɔI	ju
L1964*	ɑ	ɑ	—	—	ɔ	—
HF1962†	æ/ʌ/a		ɛ	ɔ	ɔ	I/i

*Lehiste 1964
† HF1962 = Holbrook and Fairbanks 1962.
The "/" indicates alternatives cited in study; the alternatives are the same for /aU,aI/ in HF1962. See text for details on these and other studies.

Table II. Closest monophthongal vowel in comparison with final portion of diphthong

Study	Diphthong					
	aU	aI	eI	oU	ɔI	ju
L1964*	ɔ	I	I/i	—	I	—
HF1962†	ɔ	ɛ	I	ɔ	I	u
W1965‡	U/u/o	I/i	I/i	U/u/o	I/i	—

*L1964= Lehiste 1964
† HF1962= Holbrook and Fairbanks 1962
‡ W1965 = Wise 1965
The "/" indicates alternatives cited in study. See text for details on studies.

Table III. Speech Material

Carrier Sentence:

I will say kay _____ again. (Stress on test item)

I will say KAY _____ again. (Stress on KAY)

Test Words:

Context:	b ___ d	h ___ d	IPA symbol
	bowed	howed	aU
	bide	hide	aI
	bade	hayed	eI
	bode	hoed	oU
	Boyd	Hoyd	ɔI
	beaut	hewed	ju

Four randomized lists of six repetitions of each word

List 1: Slow rate -- stress on test item
List 2: Slow rate -- stress on KAY
List 3: Fast rate -- stress on test item
List 4: Fast rate -- stress on KAY

Table IV. Mean duration and standard errors in ms for sentences. N= 192 for cells and 384 for column and row means.

Stress Condition		Rate: Slow	Rate: Fast	Row mean
Stressed	$\bar{x}$	1636	1176	1406
	$\sigma_{\bar{x}}$	19.59	16.04	17.27
Unstressed	$\bar{x}$	1637	1273	1455
	$\sigma_{\bar{x}}$	10.26	5.65	10.98
Column Mean	$\bar{x}$	1636	1224	1430
	$\sigma_{\bar{x}}$	11.04	8.84	10.26

Table V. Mean durations and standard errors in ms for target syllables pooled across diphthong. N = 96 for cells and 192 for column and row means.

		Rate		
Stressed		Slow	Fast	Row mean
[b _ d]	$\bar{x}$	387	298	342
	$\sigma_{\bar{x}}$	7.50	5.49	5.65
[h _ d]	$\bar{x}$	354	261	308
	$\sigma_{\bar{x}}$	6.69	4.72	5.30
Column Mean	$\bar{x}$	371	279	325
	$\sigma_{\bar{x}}$	5.15	3.84	3.97
Unstressed				
[b _ d]	$\bar{x}$	272	220	247
	$\sigma_{\bar{x}}$	4.24	3.11	3.25
[h _ d]	$\bar{x}$	260	199	229
	$\sigma_{\bar{x}}$	3.84	3.46	3.39
Column Mean	$\bar{x}$	266	209	238
	$\sigma_{\bar{x}}$	2.89	2.44	2.38
Grand Mean	$\bar{x}$	318	244	281
	$\sigma_{\bar{x}}$	3.98	2.89	2.80

Table VI. Mean durations and standard errors in ms for each diphthong under each condition of stress and rate in context [b _ d]. N= 16 for cells, 64 for rows and 96 for columns.

Diphthong		Rate/stress condition				
		SS	SU	FS	FU	Row mean
aU	$\bar{x}$	263	176	191	138	192
	$\sigma_{\bar{x}}$	15.31	9.33	9.67	5.23	7.66
aI	$\bar{x}$	252	164	174	125	179
	$\sigma_{\bar{x}}$	15.51	12.85	7.76	4.88	7.93
eI	$\bar{x}$	220	153	171	116	165
	$\sigma_{\bar{x}}$	14.37	8.86	7.89	3.38	6.61
oU	$\bar{x}$	220	144	163	113	160
	$\sigma_{\bar{x}}$	12.53	9.41	7.03	4.35	6.53
ɔI	$\bar{x}$	244	168	197	129	185
	$\sigma_{\bar{x}}$	14.09	10.73	8.66	3.13	7.20
ju	$\bar{x}$	171	127	139	103	135
	$\sigma_{\bar{x}}$	9.96	7.23	7.15	4.27	4.76
Column mean	$\bar{x}$	228	155	173	121	169
	$\sigma_{\bar{x}}$	6.27	4.28	3.75	2.06	2.94

Key: SS = Slow-stressed SU = Slow-unstressed
FS = Fast-stressed FU = Fast-unstressed

Table VII. Mean durations and standard errors in ms for each diphthong under each condition of stress and rate in context [h _ d]. N= 16 for cells, 64 for rows and 96 for columns

Diphthong		Rate/stress condition				
		SS	SU	FS	FU	Row mean
aU	$\bar{x}$	207	149	148	119	156
	$\sigma_{\bar{x}}$	10.59	8.58	3.78	4.20	5.39
aI	$\bar{x}$	203	150	139	116	152
	$\sigma_{\bar{x}}$	10.51	10.98	5.75	4.70	5.77
eI	$\bar{x}$	192	138	138	115	146
	$\sigma_{\bar{x}}$	10.38	6.42	3.92	4.28	4.86
oU	$\bar{x}$	182	132	143	101	140
	$\sigma_{\bar{x}}$	8.49	8.75	4.59	3.77	4.95
ɔI	$\bar{x}$	209	161	159	128	164
	$\sigma_{\bar{x}}$	11.19	9.43	6.32	4.11	5.43
ju	$\bar{x}$	179	131	138	110	140
	$\sigma_{\bar{x}}$	10.14	9.20	6.16	4.75	4.96
Column mean	$\bar{x}$	195	143	144	115	149
	$\sigma_{\bar{x}}$	4.25	3.74	2.20	1.92	2.17

Key: SS = Slow-stressed SU = Slow-unstressed
FS = Fast-stressed FU = Fast-unstressed

Table VIII. Mean distances from neutral point for diphthong onsets in each rate-stress condition (in log units).

Diphthong	Rate/stress condition			
	SS	SU	FS	FU
aU	.284	.226	.283	.250
aI	.344	.312	.348	.281
eI	.242	.229	.234	.259
oU	.245	.173	.229	.190
ɔI	.360	.306	.317	.268
ju	.505	.478	.498	.459

Key: SS = Slow-stressed SU = Slow-unstressed
FS = Fast-unstressed FU = Fast-unstressed

Table IX. Statistical comparisons of mean distance from neutral point (see text).

Pairwise t- tests

Compared	DF	Mean difference	t	p(2-t)
SS - SU	5	.043	4.895	.0045
SS - FS	5	.012	1.735	.1432
SS - FU	5	.045	3.068	.0278
SU - FS	5	.032	3.486	.0175
SU - FU	5	.002	.1800	.8645
FS - FU	5	.034	2.649	.0455

Key: SS = Slow-stressed SU = Slow-unstressed
FS = Fast-stressed FU = Fast-unstressed

Table X. Mean distances (in log units) traversed by each diphthong.

Diphthong	SS	FS	SU	FU
aU	.268	.218	.157	.120
aI	.574	.514	.465	.379
eI	.205	.179	.159	.113
oU	.206	.165	.092	.124
ɔI	.621	.529	.528	.440
ju	.255	.205	.133	.113

Key: SS = Slow-stressed
FS = Fast-stressed
SU = Slow-unstressed
FU = Fast-unstressed

Table XI. Statistical comparisons of traversed distances (in log units)

Paired t-tests

Compared	DF	Mean difference	t	p(2-t)
SS - SU	5	.099	8.762	.0003
SS - FS	5	.053	5.870	.0020
SS - FU	5	.140	7.504	.0007
SU - FS	5	.046	3.824	.0123
SU - FU	5	.041	2.232	.0760
FS - FU	5	.087	6.719	.0011

Key: SS = Slow-stressed SU = Slow-unstressed
FS = Fast-stressed FU = Fast-unstressed

Table XII. Mean angles in degrees (and standard deviation) for individual diphthongs in different rate-stress groups

Diphthong	SS	FS	SU	FU	Grand mean
aU	288 (33.8)	279 (44.5)	266 (51.2)	261 (72.8)	274 (53.0)
aI	356 (8.1)	355 (10.0)	356 (11.7)	356 (14.3)	356 (11.1)
eI	337 (40.1)	336 (36.1)	344 (43.6)	327 (65.0)	336 (47.3)
oU	333 (42.0)	322 (63.1)	348 (53.7)	319 (110.8)	331 (72.3)
ɔI	21 (6.8)	24 (11.4)	26 (9.4)	29 (5.2)	25 (8.9)
ju	197 (14.3)	199 (19.2)	198 (34.7)	190 (18.2)	196 (23.0)

Key: SS = Slow-stressed
FS = Fast-stressed
SU = Slow-unstressed
FU = Fast-unstressed

Table XIII. Percent correct classifications and probablilities using linear discriminant analysis (with resubstitution method)

Condition	SS	SU	FS	FU	Across
Parameters					
x', y', d, θ	98.96	96.35	98.95	94.27	95.05
(APP)	(.98)	(.96)	(.98)	(.92)	(.93)
x', d, θ	92.71	89.58	90.63	81.25	86.20
(APP)	(.87)	(.83)	(.86)	(.74)	(.80)
y', d, θ	95.83	94.79	92.71	88.54	89.32
(APP)	(.90)	(.92)	(.87)	(.82)	(.85)
x', y', θ	93.75	92.19	93.75	84.90	89.97
(APP)	(.91)	(.90)	(.91)	(.82)	(.87)
x', y', d	95.31	92.7	98.43	86.98	90.36
(APP)	(.92)	(.92)	(.96)	(.85)	(.88)

Key: SS = Slow-stressed SU = Slow-unstressed Across = across conditions
FS = Fast-stressed FU = Fast-unstressed

Table XIV. Percent correct classifications and probablilities using linear discriminant analysis (classification matrix for slow-stressed condition is used to classify other conditions)

Condition / Parameters	SU	FS	FU
x', y', d , θ	94.79	98.96	84.90
(APP)	(.94)	(.97)	(.84)
x', d, θ	85.42	86.98	72.92
(APP)	(.80)	(.82)	(.71)
y', d, θ	91.67	91.15	78.13
(APP)	(.89)	(.87)	(.77)
x', y', θ	91.67	95.31	85.42
(APP)	(.89)	(.91)	(.83)
x', y', d	85.94	90.63	70.83
(APP)	(.83)	(.89)	(.70)

Key: SU = Slow-unstressed FS = Fast-stressed
FU = Fast-unstressed

Table AI. Expected F_1/F_2 movements from [b] or [h] into "nucleus" of diphthongs

FROM\TO	a		ɛ		o		j	
	F1	F2	F1	F2	F1	F2	F1	F2
b	UP	LEV DOWN	UP	LEV UP–	LEV DOWN	UP	LEV	UP
h								
unanticipated	UP	DOWN	UP+	DOWN–	LEV	LEV	UP–	DOWN–
anticipated	LEV	LEV	LEV	LEV	LEV	LEV	LEV	LEV

(LEV. = level, "–" = slight, "+" = marked)

Listing of figures and captions

Figure 1.Sensory path for male production of /ɔI/ in null context. Triangles are placed every 50 ms along path indicating the extent of spectral change over the period and the direction of the path.

Figure 2. top panel: Formant pattern for female production of stressed /aU/ at slow tempo in context [h_d]. Vertical lines indicate the determined onset and offset of diphthong.
middle panel: Corresponding formant pattern after smoothing with first-order resonator with a cut-off frequency equal to 20 Hz.
Bottom Panel: Corresponding sensory path with crosses placed at diphthong onset and offset, line connecting onset and offset, and angle (see text for details)

Figure 3. Mean x'/y' values for onset of diphthongs. The bars indicate +/- 1 standard deviation from mean values.

Figure 4. Ranges of angular values for diphthongs (25th to 75th percentile).

Figure 5. left panel: Formant pattern for female production of stressed /aI/ at slow tempo in context [h_d]. Vertical lines indicate the determined onset and offset of diphthong.
right panel: Corresponding sensory path with crosses placed at diphthong onset and offset. The angle computed for this token was 347.68 degrees.

Figure 6. left panel: Formant pattern for male production of unstressed /eI/ at slow tempo in context [h_d]. Vertical lines indicate the determined onset and offset of diphthong.
right panel: Corresponding sensory with crosses placed at diphthong onset and offset. The angle computed for this token was 327.19 degrees.

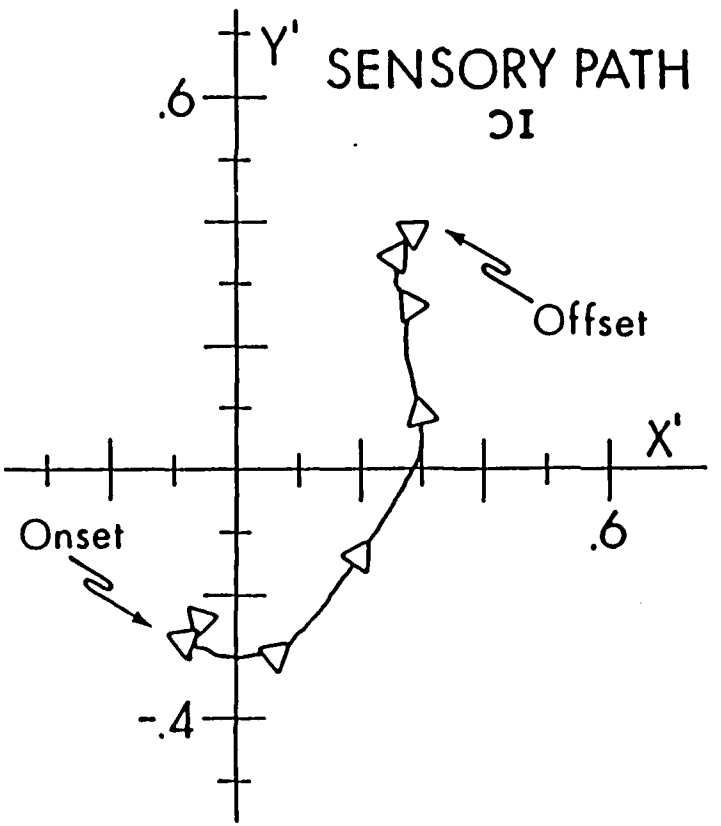
Figure 7. left panel: Formant pattern for male production of unstressed /ɔI/ at fast tempo in context [b_d]. Vertical lines indicte the determined onset and offset of diphthong.
right panel: Corresponding sensory path with crosses placed at diphthong onset and offset. The angle computed for this token was 30.75 degrees.

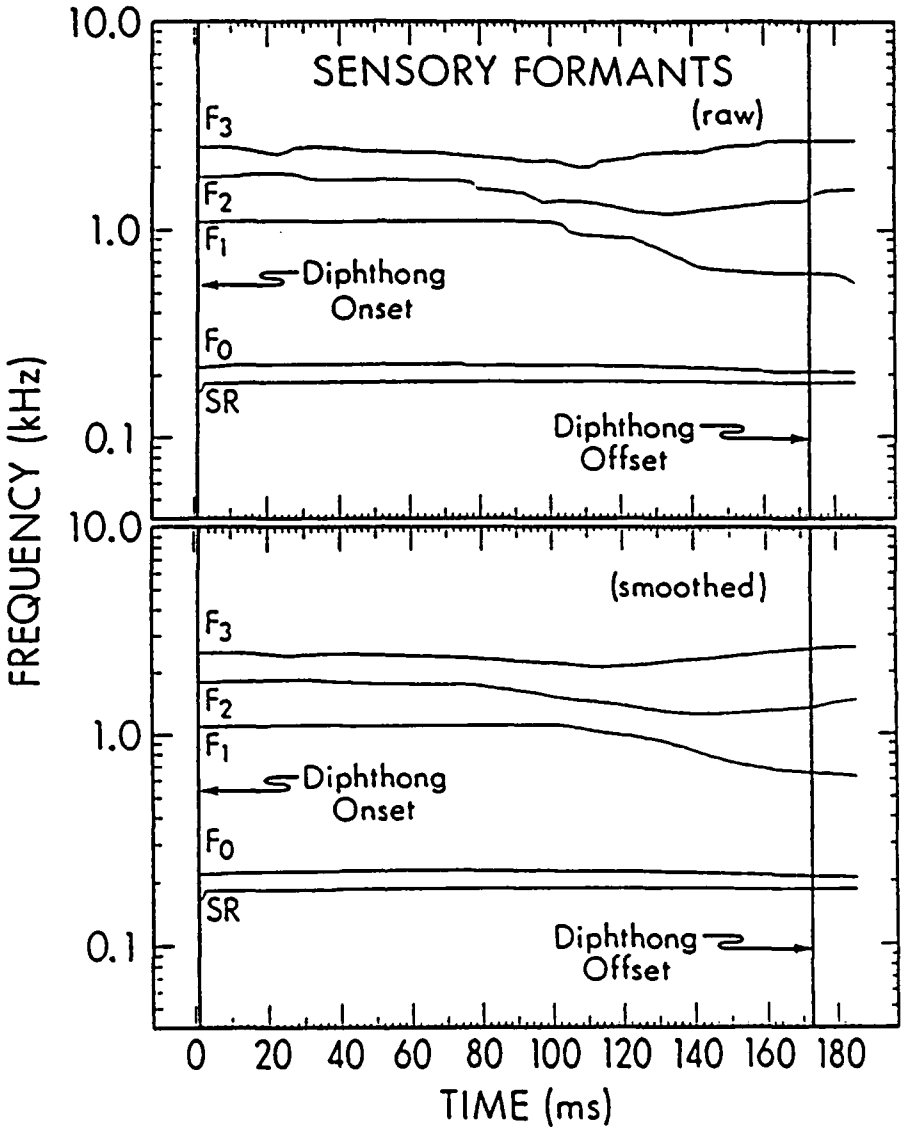
Figure 8. left panel: Formant pattern for female production of stressed /oU/ at fast tempo in context [b_d]. Vertical lines indicate the determined onset and offset of diphthong.
right panel: Corresponding sensory path with crosses placed at diphthong onset and offset. The angle computed for this token was 334.36 degrees.

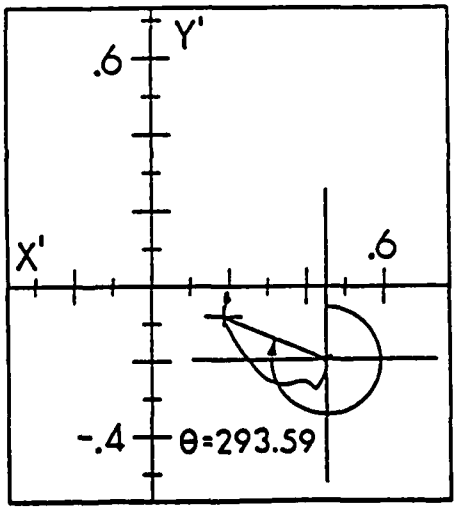
Figure 9. left panel: Formant pattern for male production of unstressed /ju/ at slow tempo in context [h_d]. Vertical lines indicate the determined onset and offset of diphthong.

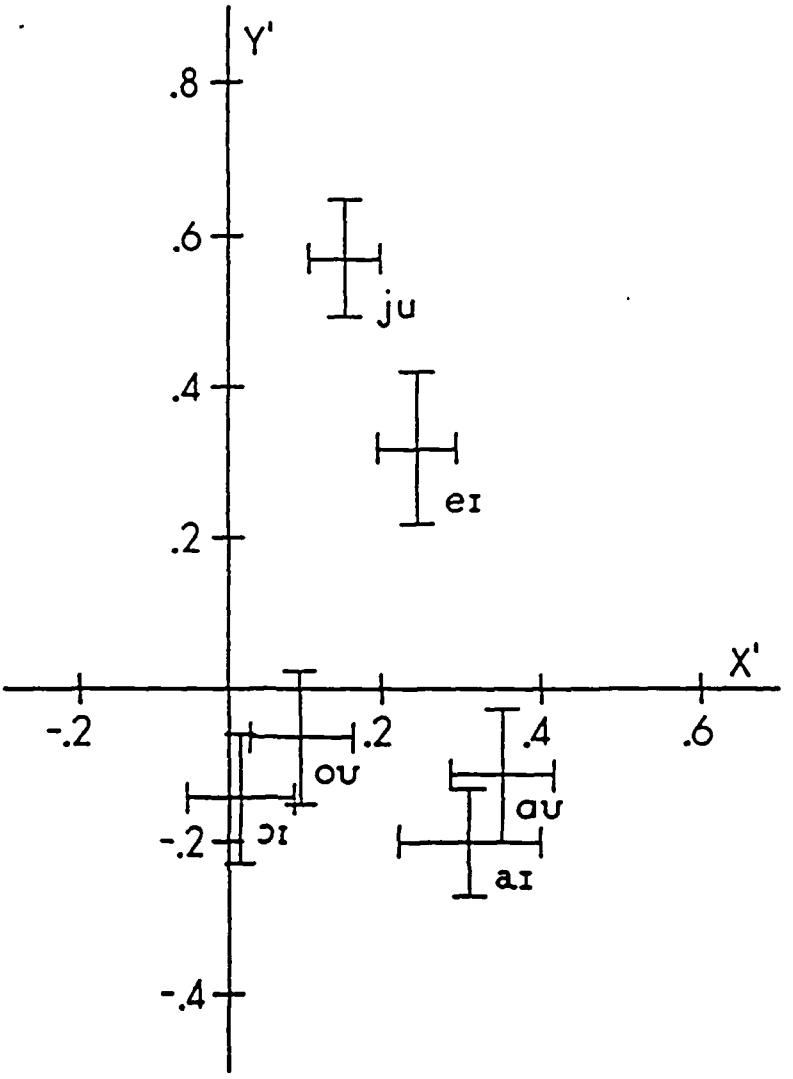
right panel: Corresponding sensory path with crosses placed at diphthong onset and offset. The angle computed for this token was 175.29 degrees.

Figure A1. Formant pattern for male production of unstressed /aI/ at slow tempo in context [b_d]. The dashed line indicates the initial determination of diphthong onset; the following vertical line indicates the revised onset.

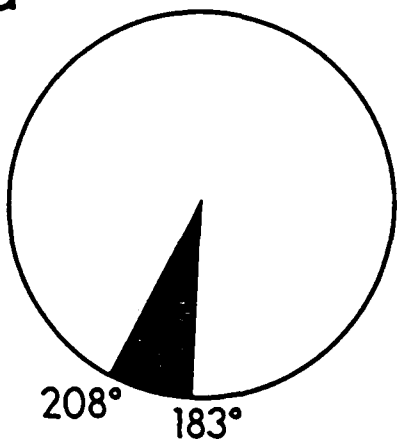

Y'
SENSORY PATH
.6
Offset
X'
Onset
.6
-.4

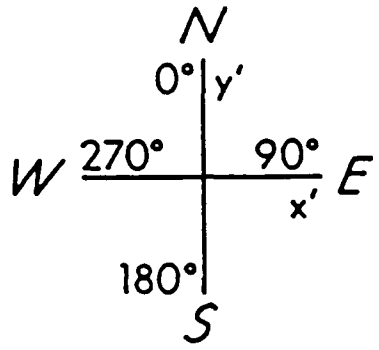

SENSORY FORMANTS
(raw)
(smoothed)
F3
F2
F1
F0
SR
Diphthong Onset
Diphthong Offset
10.0
1.0
0.1
FREQUENCY (kHz)
0 20 40 60 80 100 120 140 160 180
TIME (ms)


Y'
.6
X'
.6
-.4
θ=293.59

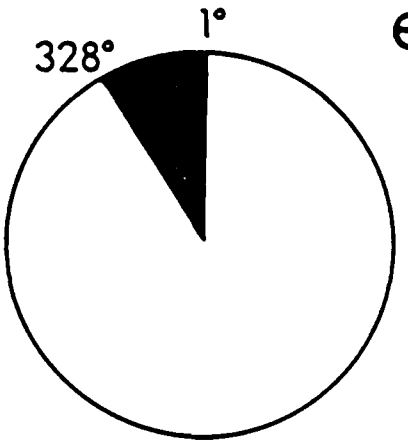

Y'
.8
.6
.4
.2
-.2
-.4
X'
-.2
.2
.4
.6
ju
eɪ
oʊ
ɔɪ
aʊ
aɪ

ju

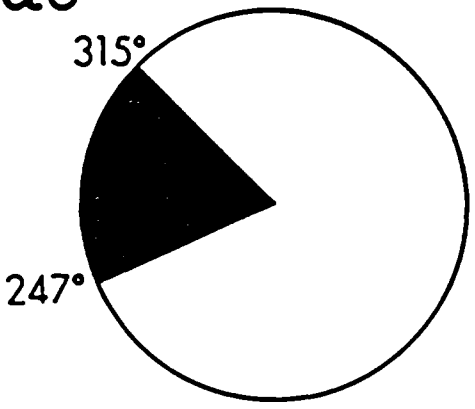




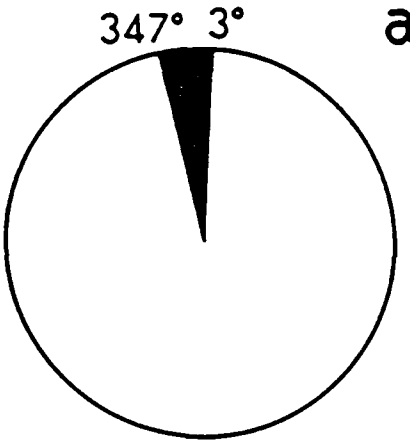
eɪ



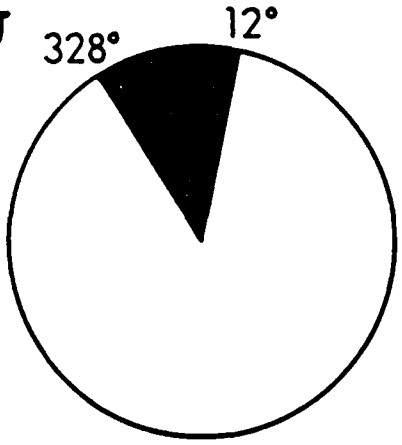
au



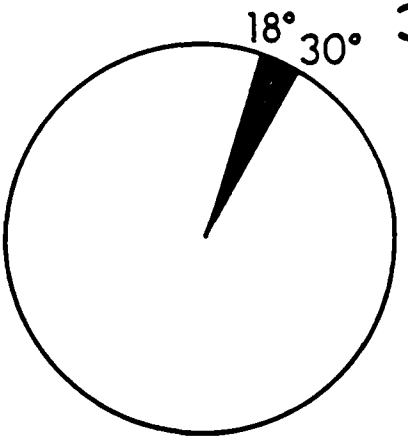
aɪ



ou



ɔɪ



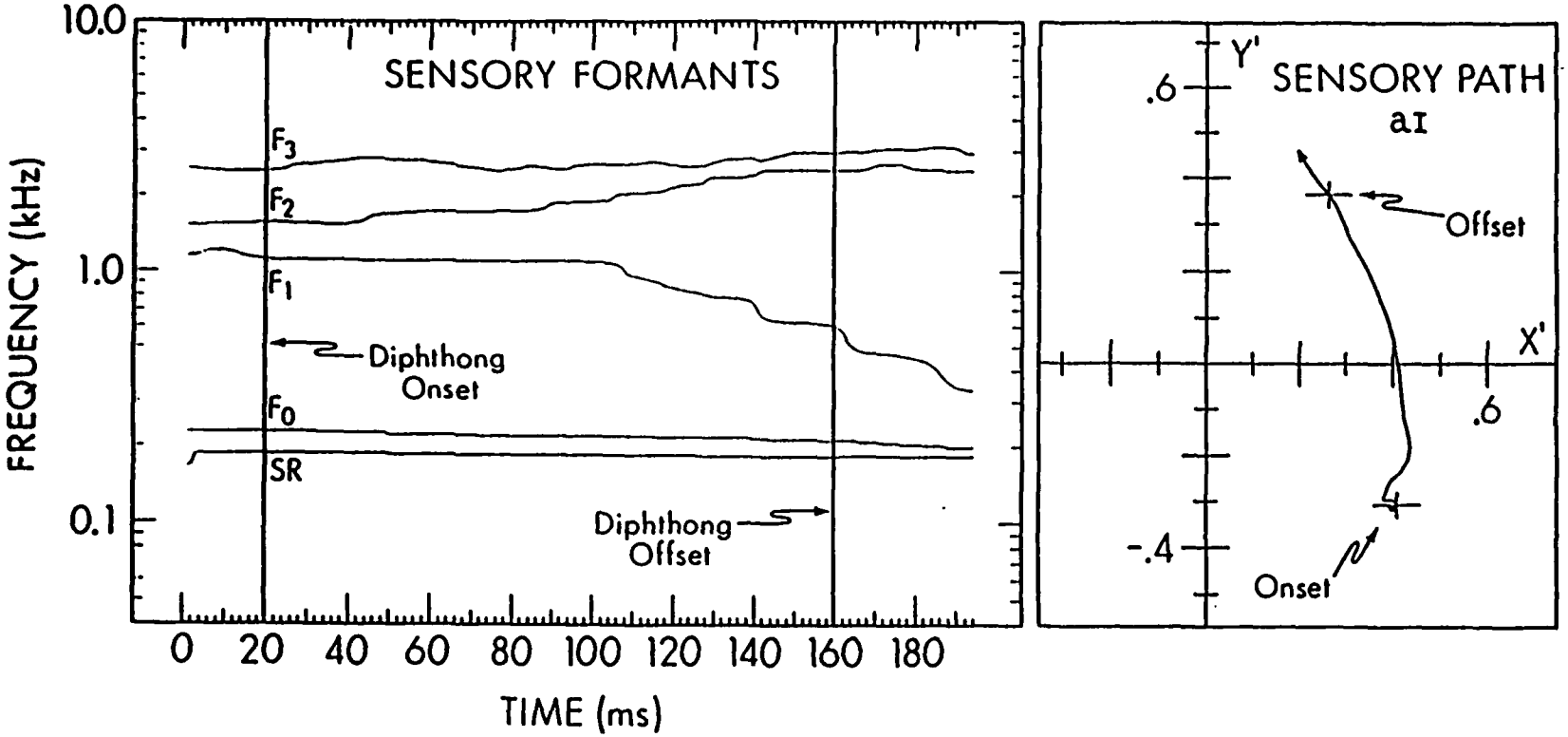
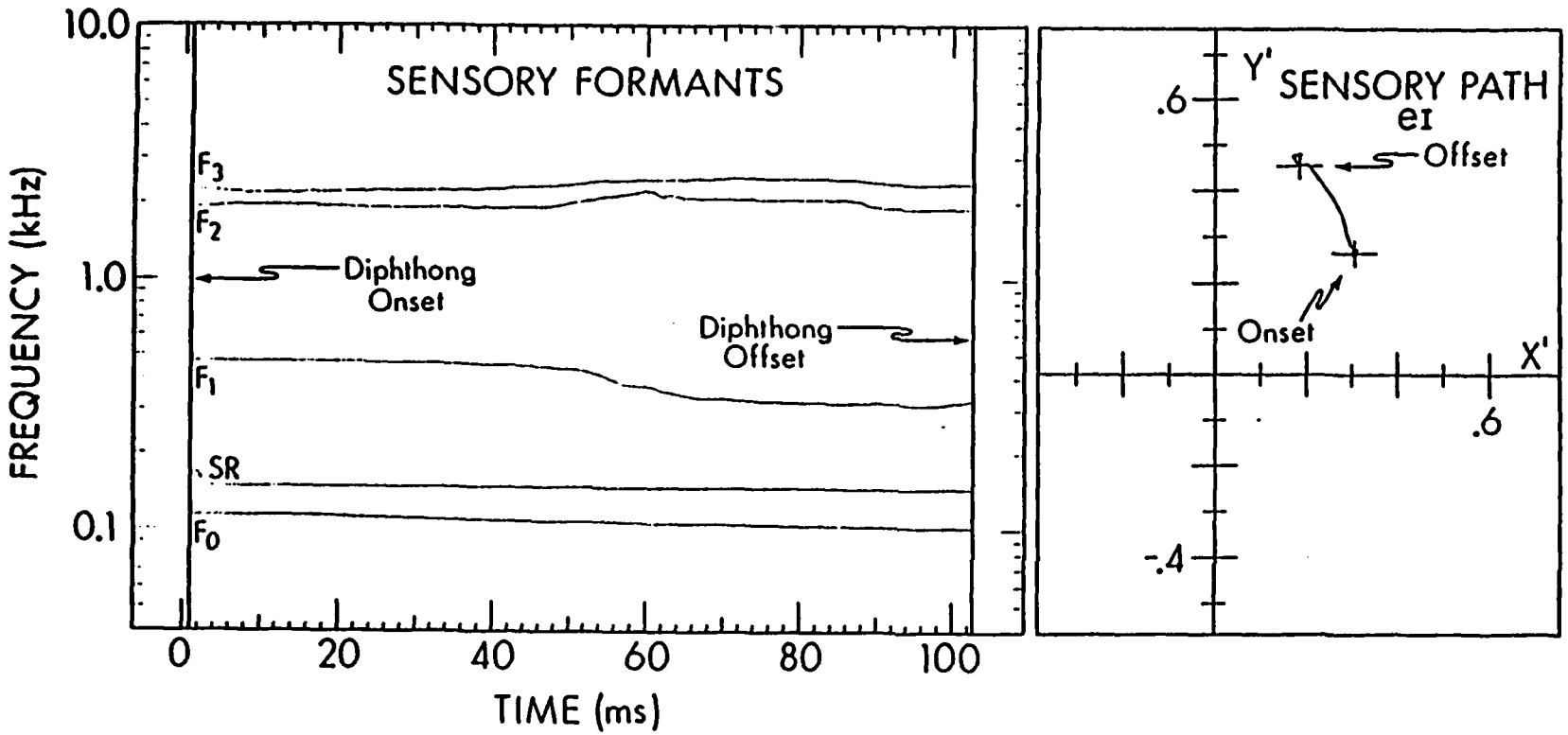
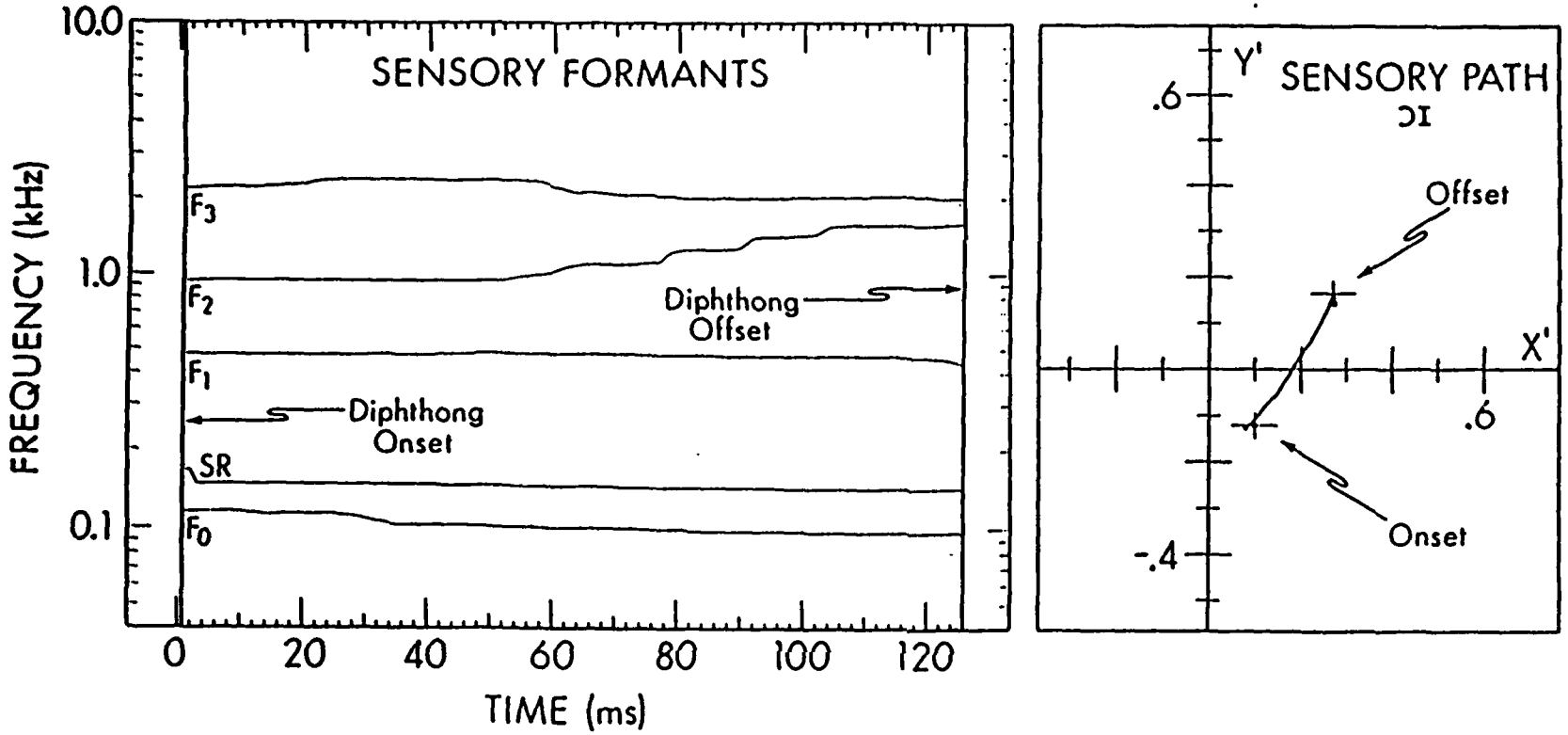

SENSORY FORMANTS
10.0
1.0
0.1
FREQUENCY (kHz)
F3
F2
F1
F0
SR
Diphthong Onset
Diphthong Offset
0 20 40 60 80 100 120 140 160 180
TIME (ms)
SENSORY PATH
aɪ
Y'
X'
.6
-.4
.6
Offset
Onset

Figure 5


SENSORY FORMANTS
10.0
1.0
0.1
FREQUENCY (kHz)
F3
F2
F1
F0
SR
Diphthong Onset
Diphthong Offset
0
20
40
60
80
100
TIME (ms)
Y' SENSORY PATH
eɪ
.6
-.4
Offset
Onset
X'
.6


SENSORY FORMANTS
FREQUENCY (kHz)
10.0
1.0
0.1
F3
F2
F1
SR
F0
Diphthong Offset
Diphthong Onset
0
20
40
60
80
100
120
TIME (ms)
SENSORY PATH
ɔI
Y'
X'
.6
-.4
.6
Offset
Onset

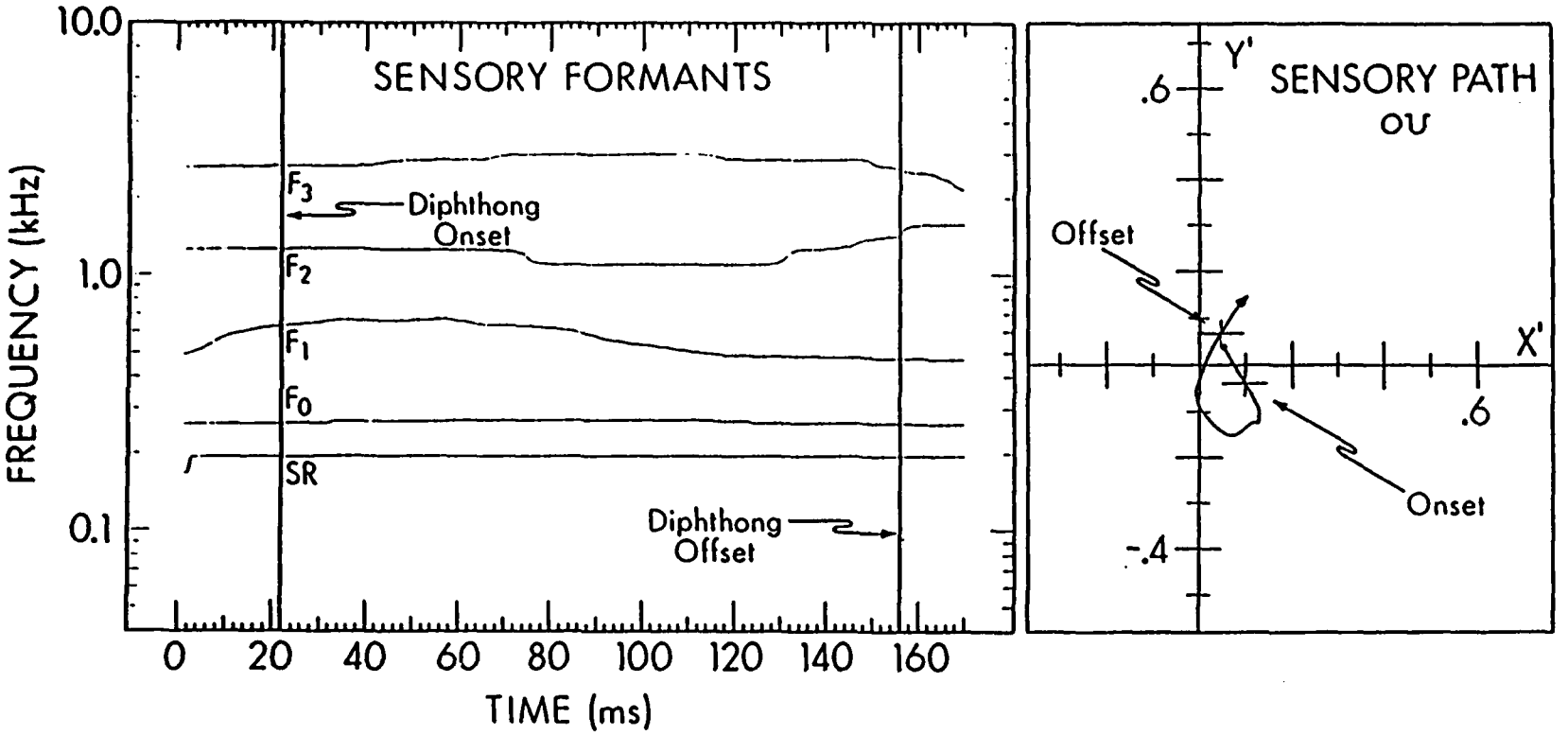
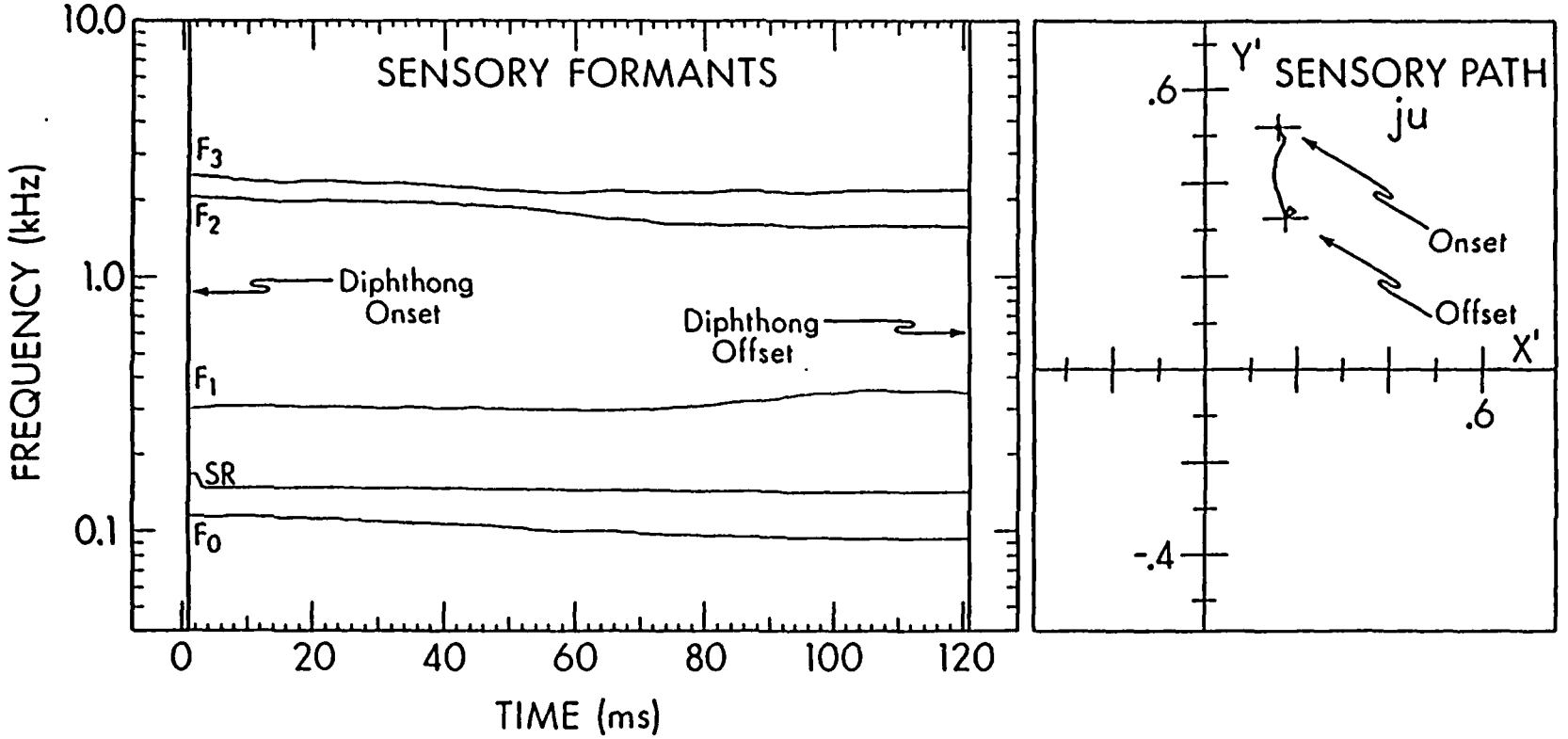

SENSORY FORMANTS
FREQUENCY (kHz)
10.0
1.0
0.1
F3
Diphthong Onset
F2
F1
F0
SR
Diphthong Offset
0
20
40
60
80
100
120
140
160
TIME (ms)
SENSORY PATH
ou
Y'
.6
X'
.6
-.4
Offset
Onset

Fig. 8



Comment Fig 9

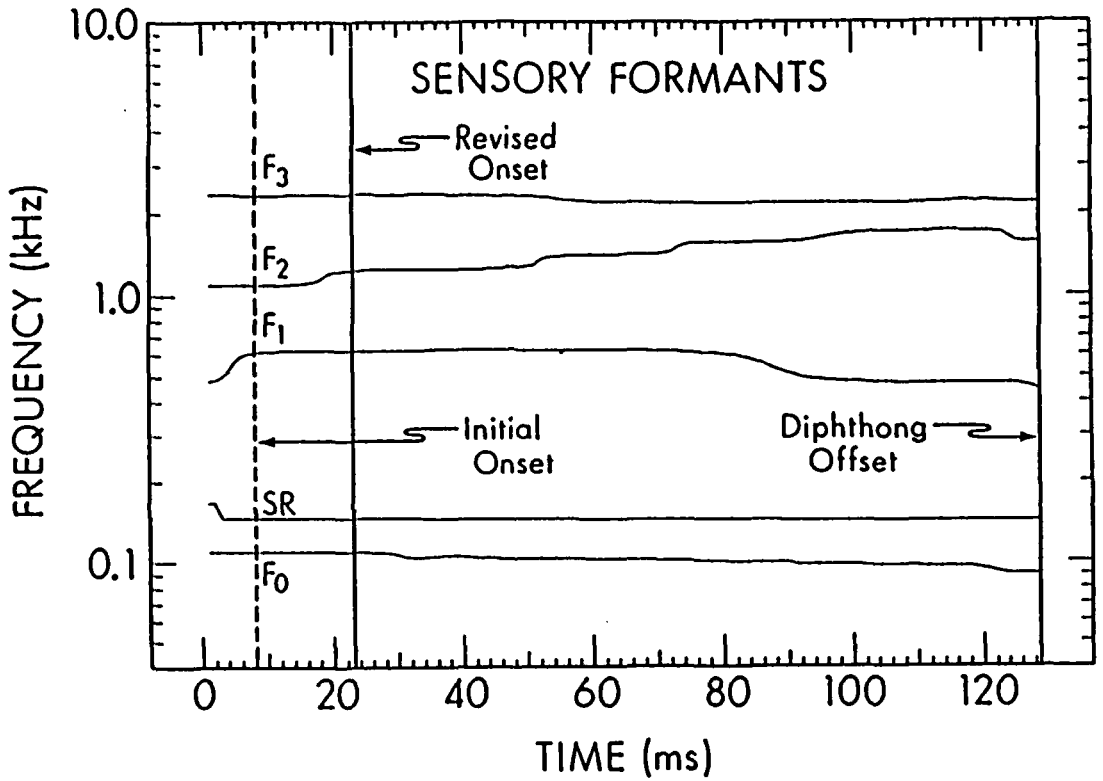

SENSORY FORMANTS
Revised Onset
Initial Onset
Diphthong Offset
F3
F2
F1
SR
F0
10.0
1.0
0.1
FREQUENCY (kHz)
0
20
40
60
80
100
120
TIME (ms)

Fig. A1