AD-A033 628 HASKINS LABS INC NEW HAVEN CONN F/G 17/2
ACOUSTIC CUES IN NATURAL SPEECH: THEIR NATURE AND POTENTIAL USE--ETC(U)
NOV 76 F S COOPER N00014-67-A-0129-0002
UNCLASSIFIED NL

1 OF 1
AD A033628

END

DATE FILMED
2-77

12

ACOUSTIC CUES IN NATURAL SPEECH: THEIR NATURE AND POTENTIAL USES IN SPEECH RECOGNITION

Final Report

ONR Contract N00014-67-A-0129-0002

ONR Contract N00014-76-C-0591

Covering the Period March 1, 1973 to November 30, 1976

Final rept. 1 Mar 73-30 Nov 76

Haskins Laboratories, Inc.
270 Crown Street
New Haven, Connecticut 06511

Nov 1976

16 p.

Franklin S. Cooper
Principal Investigator

406 643

6 pg

ACOUSTIC CUES IN NATURAL SPEECH THEIR NATURE AND POTENTIAL USES IN SPEECH RECOGNITION

Introduction

This report summarizes the work carried out by Haskins Laboratories between the years 1973 to 1976 on a research contract to study those acoustic cues in natural speech that are of potential use for speech-recognition purposes.

Objectives - The research carried out under this program had two principal objectives.

(a) To carry out basic reseach on automatically identifying the acoustic cues of natural speech. This research was aimed to bear directly on the scientific problems encountered in building a Speech Understanding System.

(b) To work closely with the principal contractors in the program whose tasks were to build complete speech-understanding systems. It was expected that this close cooperation would allow the research results to be quickly incorporated in these systems and their usefulness to be readily evaluated.

Summary of the Program's Scope and Accomplishments

A practical speech understanding system must be capable of converting the spoken message into a linguistic representation from which, after consulting the appropriate base of stored information, a useful response can be formulated. The basic structure of the spoken message is composed of phonetic elements and the very first step toward deriving a linguistic representation must be to identify the phonetic message. However, accurate automatic phonetic identification solely on the basis of acoustic data cannot be achieved by present techniques and, moreover, there is evidence that even human listeners experience difficulties under similar conditions. During most speech exchanges, higher-order information of a lexical, syntactic semantic and pragmatic nature playsa significant role in human listener's abilities to interpret the phonetic content of speech. The primary factor that distinguishes speech understanding systems from automatic isolated-word recognizers is that they generally include algorithms (called components) that generate and evaluate hypotheses at a variety of levels. These levels can often be identified as being lexical, syntactic, semantic and pragmatic as well as acoustic-phonetic.

The acoustic-phonetic component must convert the acoustic signal into a series of hypothesized phonetic strings, each with assigned likelihoods, that represents the results of a local acoustic analysis of the utterance. The system's performance at the acoustic level may be augmented by an acoustic

verification component whose responsibility it is to evaluate the acoustic evidence in favor of specific word or syllable-hypotheses that are formulated at higher levels in the system hierarchy. Similar procedures of analysis and verification are often applied at these higher levels, using evidence that must, in the first instance, be extracted by the acoustic analyzer but is sifted and evaluated in the light of the system's current knowledge of the topic being discussed. However, the scope of the work to be reported here was concerned exclusively with developing improved procedures for primary acoustic analysis and verification that could serve systems having the capabilities of performing higher levels of analysis.

Our research began with an exploration of the difficulties encountered by humans in analyzing the patterns of connected speech presented in the non-acoustic medium of a spectrogram. The task involved the analysis of an unknown sentence into words which were matched against those selected from a lexicon of spectrograms. The work was conducted in the absence of any semantic or syntactic knowledge and therefore engaged the analyst solely at the level of acoustic features. Experience gained in this experiment underlined the difficulties of extracting reliable acoustic cues, particularly from the unstressed parts of sentences.

Following such early studies, a strategy was developed for organizing the various stages of acoustic analysis that would be responsible for segmenting and labeling sections of the acoustic signal in an ordered hierarchy. The sequential arrangement featured the more reliable analyses first so that if errors were encountered in later analyses, only a limited amount of backtracking would be necessary. Furthermore, since the acoustic cues in stressed syllables are generally more sharply defined than in unstressed words and boundaries cannot be directly identified in the speech signal, an algorithm was developed to segment the signal into syllable-sized units. The output units of this segmentation algorithm were subsequently used in an exploratory study of methods for the automatic detection of relative prominence of syllable sequences.

Algorithms were also developed for several steps involved in the detailed analysis of syllabic units into constituent segments. In particular, our work included a detailed study aimed at detecting likely transition points to or from nasal consonants and characterizing the transition regions as manifestations of the onset or termination of nasality. For the future, however, it is apparent that more detailed studies of other acoustic features are required to guide the construction of decision rules for extracting other phonetic segments.

In additional work, we anticipated the need to retrieve syllables from a lexicon with the aid of only an incomplete phonetic representation. Our pilot study explored an acoustic-verification technique that compared various distance measures for their effectiveness in distinguishing monosyllabic words spoken by four speakers despite interspeaker differences.

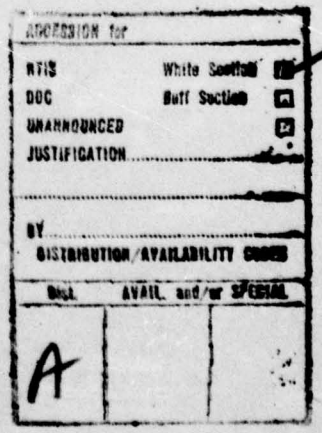
The Laboratories' syllabic segmentation algorithm was supplied to the

Systems Development Corporation and utilized in their speech-understanding system. However, the strategy that we proposed for phonetic analysis was not evaluated within any of the major contractors' systems primarily due to manpower limitations. Our strategy was based in part on the already available results of human speech perception studies and was intended to be generalizable to very large vocabularies without significant difficulties. This latter objective differed from that pursued by the major systems builders who based their systems on vocabularies of very limited size and, unfortunately, this prevented our phonetic analysis strategy from being fully implemented by any of the systems builders. Moreover, since we suffered from a shortage of manpower, we were also prevented from implementing the complete analysis ourselves before the program expired. Nevertheless, several individual components of our syllable-based speech analysis were evaluated and yielded results comparable or significantly better than those attained by previously used methods.

In addition to the work on acoustic analysis, the Laboratories also carried out development work on two major research facilities; the Digital Pattern Playback--a tool for analyzing and resynthesizing speech signals--and a software interface to the ARPA network designed for the PDP-11/45 computer and RSX-11D operating system.

To sum up, the primary focus of our work was to identify problems at the acoustic level that automatic speech recognition systems must overcome and to find promising methods for their solution that could be used by the major systems builders. Although automatic speech recognition for large vocabularies is still an unsolved problem, we can look back on the work completed under this program with the firm conviction that we now have a better understanding of the basic problems and have made some important approaches toward their eventual solution.

This report includes contributions by the following members of the research and technical staff of Haskins Laboratories: F. S. Cooper, Principal Investigator and Staff Members P. W. Nye, P.Mermelstein, J. Gaitenby, G. M. Kuhn, R. M. McGuire, L. Reiss and T. Montlick. A perusal of the individual research reports may not give the reader an adequate view of our overall effort. Hence, to overcome this shortcoming in the record, we discuss briefly in what follows the individual research efforts carried out under this grant, attempt to integrate them into the overall goals and review the conclusions drawn. In each case a detailed report on the work has been written and is attached or cited for the information of readers interested in the experimental details.



RESEARCH STUDIES PERFORMED AT HASKINS LABORATORIES

1. Speech Recognition through Spectrogram Matching

Human listeners process and identify speech signals with an ease (borne of much experience) that belies the complexity of the task. The true nature of the complexities becomes apparent quite quickly though when the listener is denied the use of his ears and is obliged to examine some nonacoustic representation of the speech signal to uncover the message. Not only do the problems come more clearly into focus, but alternative strategies for solving the problems can be explored with a view to subsequently applying these strategies in computer algorithms.

Two separate studies involving the use of spectrograms as the nonacoustic medium were carried out under the contract. The first (Ingemann and Mermelstein, 1975) employed conventional paper spectrograms of a sentence occupying several seconds in length and a lexicon of about 100 reference words. Experience soon showed that the paper shuffling task involved in organizing a large number of reference spectrograms could easily become unmanageable. A second computer-assisted study (Nye, Cooper and Mermelstein, 1975) was then carried out. Both studies aimed to:

(a) Assess the performance of humans in matching spectrograms of words in sentences with spectrograms of the same words when spoken in a reference context.

(b) Study any improvements in analysis that take place as a result of supplying to the human analyzer with feedback spectrograms (i.e., spectrograms of spoken versions of the analyzer's chosen word-sequence representing his hypotheses for the constituents of the unknown sentence frame). In speech recognition systems, such verifying feedback data could be generated by existing speech synthesis techniques.

The study led to three principal conclusions:

(i) Even when the number of words correctly matched is low, the number of syllables in the hypothesized word-sequence generally agrees with that in the unknown sentence.

(ii) The recognition of monosyllabic words is made significantly more difficult by the presence of a greater amount of phonetically similar words in the reference vocabulary. Furthermore, the number of correctly matched phonemes is always significantly larger than the number of syllables (or words) because the errors are generally substitutions of phonetically similar words. These are important facts to be aware of because an ability to discriminate among many phonetically similar words is an essential requirement for large-vocabulary recognition systems.

(iii) The study of the potential benefits of feedback of an hypothesized

sentence-frame was inconclusive. It appeared that feedback can be useful only if performance is already relatively high without feedback. When a significant number of errors is present in the hypothesized word sequence and the variation in context makes only a minor contribution to these errors, performance is not improved.

2. Spectrogram Reading of Vowel-Consonant-Vowel Sequences.

The studies of word matching in sentence contexts revealed that observers often have difficulty in accurately segmenting the speech signal into vocalic and consonantal segments. However, even after the vowel-consonant segmental pattern is determined, additional problems remain, such as those involved in selecting the appropriate consonant when no additional lexical information is available. Such problems are, of course, latent in the automatic recognition of continuous speech and it was therefore appropriate to conduct an exploratory study (Kuhn and McGuire, 1974). The aims of the study were to:

(a) Study the confusions among consonants that are encountered by experienced acoustic-phoneticians and then examine the differences in the acoustic cues for these consonants with the aid of spectrograms.

(b) Observe the effects of concentrated learning with feedback on improving the spectrographic identifiability of tokens having the same general phonological context.

The conclusions of our vowel-consonant-vowel (VCV) study were that:

(i) Place of production errors are the most frequently encountered among consonant errors in a VCV environment. It is known that the spectral positions of place of production cues are shifted significantly depending on the vowel environment. Manner and voicing errors occur much less frequently.

(ii) In the course of learning sessions, overall identification of consonants improved significantly from 75 to 90 percent. Stops and fricatives showed the largest improvement. Identification of nasals and semivowels proved to be more resistant to learning.

(iii) Even after concentrated learning on cues exhibited in similar contexts, a significant number of errors remained. A conventional spectrographic representation of the acoustic cues may not be adequate for perfect recognition. One may have to use additional cues not easily seen in spectrographic presentations.

3. Automatic Segmentation into Syllablic Units

Segmentation of the continuous speech signal into articulatorily,or phonologically,relevant units must be one of the first steps in any analysis procedure. Hence, following the spectrograms reading experiments, an attempt at finding a satisfactory way of segmenting the speech signal became a matter

of first priority. The research study led to the selection of the syllable as the most promising speech unit; having not only a linguistic identity but a relatively stable physical manifestation as well. Additional factors in favor of the syllable derived from the fact that the interaction between adjacent syllabic units is much less than that observed between their constituent phonetic segments. Moreover, it appeared likely that more detailed analyses could be successfully made to subsequently concentrate on their internal segmental structure. Finally, the segmentation of the speech signal into words could be achieved by mapping syllabic-sized units into an appropriately structured syllable-based lexicon.

These were the reasons that led us to explore the development of an algorithm that would automatically segment continuous speech signals into syllable sized units (Mermelstein, 1975b).

The conclusions drawn from our study were that:

(i) Syllabic units can be isolated in continuous speech by simple automatic means. The algorithm was tested on 400 syllables of continuous speech and missed only 6.9 percent of the syllables and inserted barely 2.6 percent of additional syllables relative to a nominal, slow-speech syllable count.

(ii) The syllabic boundaries that are chosen by the algorithm, however, do not generally correspond to boundaries assigned on the basis of phonological criteria.

4. A Strategy for Acoustic Analysis

Having developed an algorithm that would segment the speech signal into syllable-sized units, the next step became the development of a strategy for analyzing these units in an effort to identify the constituent phonetic segments. The study we undertook had several basic aims the rationale of which Mermelstein (1975a) described in his published report. These objectives were that:

(a) Having identified a few segments, advantage should be taken of the phonological constraints on adjacent segments thus eliminating the necessity to consider every phone as an hypothesis for each identified segment.

(b) Since segmentation and labeling generally require similar analysis operations, our strategy should be to combine the two procedures so that the signal is segmented only at points where a labeling difference is found between the adjacent segments.

(c) The hypothesized segmental constituents of syllabic units should be used to retrieve similarly represented items from a reference syllabary. Thus all words that share a stressed syllable may be readily retrieved on the basis of acoustic information from the stressed syllable alone. The hypotheses may be more or less specific. Since the general hypotheses subsume more specific

ones, too many reference forms may be retrieved in response to a particular hypothesis. In such cases, additional analyses would be performed to reduce the number of comparisons to reference items.

Our major conclusions from this study were as follows:

(i) An hierarchical organization is an effective device to exploit our current knowledge about the phonological constraints within syllabic segments. Its special advantage is that it allows easy integration of independent decision modules into the overall system.

(ii) Originally, we considered implementing the decision structure in a deterministic manner. However, since small finite error probabilities are associated even with early decisions a probabilistic structure appears to be more appropriate. Decisions at every node of the decision tree have a priori and a posteriori probability assignments. The best hypothesis is associated with the highest a posteriori probability. If the first hypothesis breaks down, branches with lower probability can be followed.

(iii) The speech signal is composed of dynamic segments. Perceptual categorization of these segments is based on numerous acoustic parameters whose detailed contribution to an individual decision are not yet known. Accurate phonemic categorization of segments requires a better understanding of the perceptual roles of the various acoustic parameters. This appears to be the strongest limitation to our acoustic phonetic decoding capability today.

5. Detecting Nasals in Continuous Speech

The building of a detection component for nasal consonants was undertaken as a step toward implementing our analysis strategy for extracting segmental information from the syllabic units. The work also set out with the aim of taking advantage of the information latent in the phonetic context of segments (Mermelstein, 1975c).

Hypotheses concerning the possible existence of nasals were formulated according to the context-dependent strategy outlined in Section 4. We first identified the spectral transition points that could mark the onset or termination of nasal-murmur segments then using the segmentation of the speech stream into syllable-sized units we took advantage of the phonological constraint that a syllable has, at most, two nasal-murmur segments; one prior to the syllabic vowel and one between the vowel and the end of the syllable. Additionally, we investigated to what extent knowledge of the direction of the transition, into or out of the nasal, was useful in attaining improved recognition of the existence of nasal segments.

This study led to the following conclusions:

(i) The transition to and from nasal segments can be effectively characterized by the time-varying characteristics of four simple acoustic measures, the relative energy change in the frequency bands 0-1, 1-2 and 2-5 kHz and the frequency centroid of the 0-500 Hz band. Using multivariate statistics on four samples of these measures spaced 12.8 msec apart, a 91 percent correct nasal/nonnasal decision rate can be attained for data of two speakers. This categorization rate is significantly better than could be attained using stationary statistics on the same or similar measures.

(ii) Careful selection of the point of maximum spectral change is critical to the success of the procedure. The usefulness of these measurements toward the separation of nasal and nonnasal categories drops rapidly as one moves away from the point of maximal spectral change.

6. Distance Measures for Speech Recognition

The most useful metric to represent the acoustic similarity of unknown syllables is one that can also predict the perceptual similarity of those syllables. The aim of this exploratory study (Mermelstein, 1976b) was to review available data from speech perception and speech transmission studies on the confusability of speech sounds and to express this confusability as a multidimensional distance in phonetic space. Some desirable properties that distance measures should possess for the accurate verification of syllable hypotheses were identified in the light of the perceptual data.

Our experimental study explored the use of a two-dimensional mel-based cepstral distance measure of the distance between many unknown syllables spoken by various speakers. Syllable templates obtained by combining information from all available productions of a given monosyllabic word could be used to maximize the similarity between an unknown token of a word and its stored template. Additionally,the significance of local cepstral differences was assessed with the aid of estimates of the variability of those measures over the set of known productions of that word.

The study concluded that:

(i) An ability to weigh observed differences according to their significance was necessary for successful verification.

(ii) Speech synthesis techniques can yield representative tokens of the hypothesized syllables which are acceptable to a listener but they do not provide information concerning the degree to which variations from the tokens may occur. Therefore stored templates augmented by variability information offer a better short-term solution to the verification problem.

7. Acoustic Determinants of Stop-consonant Place Perception

Identification of the place-of-production feature of stop-consonants is a difficult task for any speech recognition system. A study reported by Kuhn (1975) focused on a theory for perceptual place assessment based on distinguishing the front-cavity resonance (i.e., the resonant frequency of the cavity of the vocal-tract lying immediately anterior to the major constriction that produces the consonant). The study showed that, if one assumes that the front-cavity resonance can be detected by human perceptual mechanisms, one can explain certain perceptual behaviors with synthetic speech stimuli that would otherwise appear anomalous if attempts were to be made to explain them solely on the basis of acoustic considerations. Additional data on perceptual measurements are included in a Ph. D. Thesis entitled "An Experimental Study of the Acoustic Determinants of Stop Consonant Place Perception: Observations from the Synthesis of Single Formant Stimuli" to be submitted to the Department of Linguistics at the University of Connecticut by G. M. Kuhn. This research study attempted to assess the role of resonances of the front cavity of the vocal-tract in the perception of intervocalic stop consonants.

The relevance of this study to the automatic analysis of acoustic speech data was based on three hypotheses:

(a) A front cavity resonance frequency estimate can be obtained automatically from the information in the speech signal.

(b) The cues for place of articulation of consonants can be described concisely from an articulatory viewpoint and the front-cavity resonance serves as an aid to the listener in assessing the place of articulation.

(c) A front-cavity resonance frequency estimate may serve as a speaker-independent articulatory reference. Front-cavity lengths may be more similar across speakers than the lengths of their entire vocal tracts.

Kuhn concluded that:

(i) A front-cavity frequency estimate can be made from speech data by weighting the spectra by the middle and inner ear transfer functions, converting the frequency scale to mels, smoothing with filters having equal bandwidths in mels, and selecting the most prominent spectral peak.

(ii) The relative contributions of the second or third formant frequency to stop-consonant place of perception changes as the front-cavity affiliation of those formant change. The formant that contributes most to correct place perception appears to be the one that is most closely associated with the front cavity.

8. Acoustic Determinants of Perceived Prominence

This study (Gaitenby, 1976) sought to find acoustic indicators of the most prominent syllables within an unknown utterance. Since the most prominent syllables generally carry more detailed acoustic information, acoustic-based hypotheses concerning the phonetic makeup of such syllables are more likely to be correct than are the hypotheses for the less prominent syllables. A vocabulary organized around certain syllables that are likely to be prominent can then be used to retrieve hypotheses concerning the less prominent syllables. The study computed a weighted intensity function for the speech signal and compared a measure of syllable prominence derived automatically from that function with listeners' perceptual judgments of the relative prominence of the syllables. Weighted intensity, a measure roughly representing the relative loudness of speech as a function of time, was used previously for segmentation of the signal into syllable-sized units.

Our conclusions were that weighted intensity is also a reliable indicator of relative prominence among syllables. Predictions based on that measure were in substantial agreement with perceptual judgments of the same speech material.

9. A Digital Pattern Playback for the Analysis and Manipulation of Speech Signals.

The Digital Pattern Playback (Nye, et al., 1975) is a new computer-based research tool for the analysis manipulation and resynthesis of speech data. Its original design was supported by a grant from the National Science Foundation and its construction was completed under this contract. The Digital Pattern Playback,which is similar to an earlier analog pattern playback, permits the generation of artificial speech sounds containing variants of the features being studied. An important addition is the ability to display gray-scale digital spectrograms practically instantaneously after the utterance is spoken. Through a connection to a general-purpose computer, previously recorded utterances can be retrieved and compared to newly recorded speech. Furthermore, focusing on the perceptual differences resulting from small spectrographically defined changes in the acoustic signal, the perceptual effects of acoustic features can be rapidly evaluated.

This instrument has already received extensive use for the analysis of acoustic features of utterances such as voice-onset time, nasal resonances, segmental durations and formant frequency variations. It has also been extensively used in simulating automatic feature assignment to unknown utterances prior to the implementation of an automatic extraction routine that isolated those features in an exploratory speech-recognition system (see Section 1).

10. An ARPANET software interface for DEC PDP-11/45, R3X-11D.

An interface package of four programs (tasks) has been designed to operate on the Digital Equipment Corporation PDP-11/45 computer under the

RSX-11D operating system. The purpose of the undertaking was to make possible the exchange of data files and messages relating to the contract work between the Laboratories and other ARPA contractors engaged in the speech understanding program. On completion, the interface package was transmitted to about 12 other PDP-11/45 users connected to the ARPA net who had expressed the need for Network/RSX compatibility.

The software package uses a Very Distant Host (VDH) hardware interface manufactured by A Consultant. It consists of a device handler called VD (for the VDH interface) which contains the logic for the Reliable Transmission Packet protocol. A second psuedo device handler called NT implements the IMP-HOST and HOST-HOST protocols. The third component is a task called TELENET that allows the user to connect his terminal to any server HOST on the network through NT. Finally FTP is a user task that implements the network standard File Transfer Protocol. This routine allows the user to connect to a remote HOST, perform manipulations upon its file system and transfer files between the remote system and the local RSX-11D file system. At the present time only ASCII files may be transferred in this manner.

Details of the package structure and operating characteristics are described in an operators manual intended for use in the Laboratories (McGuire, 1975).

PUBLICATIONS, TALKS AND REPORTS

The following technical publications and oral papers reported on research carried out with substantial support from this contract. In most cases a copy of the technical paper or Haskins Laboratories Status Report is attached to this report.

Cooper, F. S. and P. Mermelstein. (1974) On Finding One's Way from Phonetic Text to Spoken Words and Back. *Presented at the 37th Meeting of the American Society for Information Science, Atlanta, Ga., October.*

Gaitenby, J. (1975) Stress and the Elastic Syllable: An Acoustic Method for Delineating Lexical Stress Patterns in Connected Speech. Haskins Laboratories Status Report on Speech Research, SR-41, 137-152. Presented at the 88th Meeting of the Acoustical Society of America, St. Louis, Mo., November.

Gaitenby, J. (1976) Weighted Intensity as an Acoustic Determinant of Perceived Prominence in Unknown Utterances. To appear in Haskins Laboratories Status Report on Speech Research.

Ingemann, F. and P. Mermelstein (1975) Speech Recognition through Spectrogram Matching. J. Acoust. Soc. Am. 57, 253-255. Presented at the 88th Meeting of the Acoustical Society of America, St. Louis, Mo. November 1974.

Kuhn, G. M. and R. M. McGuire (1974) Results of VCV Spectrogram-Reading Experiment. Haskins Laboratories Status Report on Speech Research, SR-39/40, 67-69.

Kuhn, G. M. (1975) On the Front Cavity Resonance and Its Possible Role in Speech Perception, J. Acoust. Soc. Am. 58, 578-585.

Mermelstein, P. (1975a) A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech, IEEE Trans. ASSP. 23, 79-82. *Presented at the IEEE Symposium Speech Recognition held at Carnegie-Mellon University, Pittsburgh, Pa., April 1974.*

Mermelstein, P. (1975b). Automatic Segmentation of Speech into Syllabic Units. J. Acoust. Soc. Am. 58, 880-883. Presented at the 87th Meeting of the Acoustical Society of America, New York, N. Y., 1974. [Also in Haskins Laboratories Status Report on Speech Research, SR-42/43, 247-256.]

Mermelstein, P. (1975c) On Detecting Nasals in Continuous Speech. Haskins Laboratories Status Report on Speech Research, SR-44, 83-94. [Also in J. Acoust. Soc. Am. 61 (in press).]

Mermelstein, P. (1976a) The Syntax of Acoustic Segments. In Conference Record, 1976 IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, Pa., April 1976, pp. 33-36.

Mermelstein, P. (1976b) Distance Measures for Speech Recognition-Psychological and Instrumental. Haskins Laboratories Status Report on Speech Research, SR-47, 91-103. Presented at the 1976 Joint Workshop on Pattern Recognition

and Artificial Intelligence, Hyannis, Mass., June 1976. [Also in Pattern Recognition and Artificial Intelligence, ed. by C. H. Chen (N.Y.: Academic Press, in press).]

Mermelstein, P. and S. Levinson. (1976) Speech Recognition-Acoustic, Phonetic and Formal-Language Models. In Proceedings of the Fourth New England Bioengineering Conference (Yale University, New Haven, Conn., May 1976), ed. by S. Saha. (N.Y.: Pergamon Press), pp. 475-477.

McGuire, R. M. (1975) RSX-11D Telnet and File Transfer Program Users Guide. Haskins Laboratories' Internal Memorandum.

Nye, P. W., F. S. Cooper and P. Mermelstein (1975). Interactive Experiments with a Digital Pattern Playback. Presented at the 90th Meeting of the Acoustical Society of America, San Francisco, Calif. November 1975. To appear in Haskins Laboratories Status Report on Speech Research.

Nye, P. W., L. J. Reiss, F.S. Cooper, R. M. McGuire, P. Mermelstein and T. Montlick. (1975) A Digital Pattern Playback for the Analysis and Manipulation of Speech Signals. Haskins Laboratories Status Report on Speech Research, SR-44, 95-107.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Haskins Laboratories, Inc. 270 Crown Street New Haven, CT 06511	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP N/A

3. REPORT TITLE

Acoustic Cues in Natural Speech: Their Nature and Potential Uses in Speech Recognition

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final Report covering the period March 1, 1973 to November 30, 1976

5. AUTHOR(S) (First name, middle initial, last name)

Staff of Haskins Laboratories; Franklin S. Cooper, P.I.

6. REPORT DATE December 10, 1976	7a. TOTAL NO. OF PAGES 13	7b. NO. OF REFS 15
8a. CONTRACT OR GRANT NO. ONR Contract N00014-67-A-0129-0002 b. ~~XXXXXXXXX~~ and ONR Contract N00014-76-C-0591 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES N/A	12. SPONSORING MILITARY ACTIVITY See No. 8

13. ABSTRACT

This final report summarizes the work carried out by Haskins Laboratories between the years 1973 to 1976 on research contracts to study those acoustic cues in natural speech that are of potential use for speech-recognition purposes. The principal objectives of the program were: (a) To carry out basic research on automatically identifying the acoustic cues of natural speech. This research was aimed to bear directly on the scientific problems encountered in building a Speech Understanding System. (b) To work closely with the principal contractors in the program whose tasks were to build complete speech-understanding systems. It was expected that this close cooperation would allow the research results to be quickly incorporated in these systems and their usefulness to be readily evaluated.

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101-807-6811

Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech spectrogram matching						
Speech spectrogram reading						
Syllabic segmentation - automatic						
Acoustic analysis - nasals						
Distance measures - acoustic						
Perception - stop consonants						
Resonance - vocal tract front cavity						
Perceived prominence - acoustic determinants						
Digital speech playback and analysis						
ARPA net software						