BOLT BERANEK AND NEWMAN INC

CONSULTING • DEVELOPMENT • RESEARCH

Report No. 2351
Job No. 11545

INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE

SEMIANNUAL TECHNICAL REPORT NO. 2, SECTION 1

TASK 1: SECOND-LANGUAGE LEARNING

15 March 1972

ARPA ORDER NO. 890, Amendment No. 6

Sponsored by the Advanced Research Projects Agency
Department of Defense, under Air Force Office of
Scientific Research Contract F44620-71-C-0065

Prepared for:

Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia 22209

**Approved for public release;
distribution unlimited.**

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)*
Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE. Task 1: SECOND-LANGUAGE LEARNING

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

5. AUTHOR(S) *(First name, middle initial, last name)*
Daniel N. Kalikow

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
15 March 1972	51	3

8a. CONTRACT OR GRANT NO.
F44620-71-C-0065

b. PROJECT NO. AO 890-6

c. 61101D

d. 681313

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
AFOSR TR 72-1321

10. DISTRIBUTION STATEMENT **Approved for public release; distribution unlimited.**

11. SUPPLEMENTARY NOTES
TECH,OTHER

12. SPONSORING MILITARY ACTIVITY
Air Force Office of Scientific Research (NL)
1400 Wilson Boulevard
Arlington, Virginia 22209

13. ABSTRACT

In this report, we outline the administrative setting and describe the experimental design to be used in field testing the Mark II model of the Automated Pronunciation Instructor (API) system. We present the draft instructional curriculum for the Spanish-English and the English-Mandarin Chinese language pairs, and we describe the hardware, pedagogical rationale, and software that have been developed to teach that curriculum.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
language learning man-computer interaction computer-aided instruction speech analysis.						

1b

Report No. 2351 Bolt Beranek and Newman Inc.

INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE

SEMIANNUAL TECHNICAL REPORT NO. 2, SECTION 1
TASK 1: SECOND-LANGUAGE LEARNING

15 March 1972

by

Daniel N. Kalikow

ARPA Order No. 890, Amendment No. 6
Sponsored by the Advanced Research Projects Agency,
Department of Defense, under Air Force Office of
Scientific Research Contract F44620-71-C-0065

Prepared for

Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia 22209

11a

TABLE OF CONTENTS

SEMIANNUAL TECHNICAL REPORT NO. 2, SECTION 1
PERIOD 1 July to 31 December 1971

ARPA Order No. 890
Program Code No. 1D20
Contractor: Bolt Beranek and Newman Inc.
Effective Date of Contract: 1 January 1971
Contract Expiration Date: 31 December 1971
Amount of Contract: $340,461
Contract No. F44620-71-C-0065
Principal Investigators: John A. Swets
Daniel N. Kalikow
Mario C. Grignetti
Duncan C. Miller
Telephone No. 617-491-1850
Title: Information Processing Models and Computer Aids for Human Performance

TASK 1: SECOND-LANGUAGE LEARNING

1. Technical Problem

The task is to carry out the final development of a computer-based system for automated instruction of the new speech sounds of second languages, and to field-test this system in the Defense Language Institute (DLI) instructional environment.

2. General Methodology

Laboratory experiments and field evaluations.

3. Technical Results

In this report, we outline the administrative setting and describe the experimental design to be used in field testing the Mark II model of the Automated Pronunciation Instructor (API) system. We present the draft instructional curriculum for the Spanish-English and the English-Mandarin Chinese language pairs, and we describe the hardware, pedagogical rationale, and software that have been developed to teach that curriculum.

4. Department of Defense Implications

Language schools of the Department of Defense give instruction in approximately 65 languages to over 200,000 students each year. The systems under development are designed to facilitate this instructional process.

ACKNOWLEDGMENTS

The author gratefully acknowledges the contributions of the other members of the project team: Douglas W. Dodds, Dennis H. Klatt, Ann M. Rollins, Kenneth N. Stevens, John A. Swets, and Thomas R. Willemain. The comments of Robert Lado, Raymond S. Nickerson, and Peter Rosenbaum are also appreciated.

1. PREFACE

The present contract is a partial continuation of a research program begun in 1966 under ARPA sponsorship. Of the four tasks eventually funded under Contract F44620-67-C-0033, with the Air Force Office of Scientific Research, the first two tasks were awarded continuing support under the present contract. Those tasks are:

1. Second-language learning
2. Models of man-computer interaction

The present Semiannual Technical report covers the progress made in the first of these tasks during the second six months of the new contract. We have bound the reports of the two tasks separately to facilitate their distribution and use. In addition to a copy of this page, both sections of this report contain an appropriate subset of the documentation data required for the whole report: a contract information page, a summary sheet for the particular task at hand, and a DD Form 1473 for document control.

2. INTRODUCTION

The objective of this research is to carry out further development of a computer-based system for automated instruction in the pronunciation of a second language, and to field test this system in the Defense Language Institute instructional environment. A prototype version of this system, called the Automated Pronunciation Instructor (API) was developed and tested under AFOSR Contract F44620-67-C-0033. The results of that program were described in that project's final report.

The present research program involves a three-year effort, the final two years of which will be primarily devoted to DLI field testing. The first year was devoted primarily to the preparation of the Mark II version of the API, and to the other scientific and administrative work that must precede the field tests. Existing software for aiding Spanish-speaking students in acquiring the speech sounds of English was expanded. Year 2 is to involve field testing of the system under DLI auspices, at a foreign military base carrying out Spanish-English language training. Year 3 is to include a field test at DLI-Monterey, where the system will be used to aid American English-speaking students in learning the sounds of Mandarin Chinese.

This report covers the first year of work on the contract. It treats the following areas of activity:

Section 3: Administrative and experimental plans
Section 4: Instructional software

During the past six-month period, we have continued our research in the phonology of Mandarin Chinese. We leave a fuller account of that work for a later report, though we will mention new developments in passing.

3. ADMINISTRATIVE AND EXPERIMENTAL PLANS

Since our last report, we have conducted extensive discussions with DLI officials and other concerned parties, with the result that the arrangements projected previously are now revised in several important respects. The most important of these is a change in the site for Year 2's field evaluation for the Spanish-English language pair. DLI stated that they could not, within the continental United States, provide adequate numbers of Spanish-speaking students (Ss) at an appropriately early stage of exposure to English. They therefore recommended that we carry out the Spanish-English (S-E) field testing within the context of an introductory intensive English training course given by the Spanish armed forces, in Spain, to Spanish military personnel, using DLI-supplied curriculum. Since this was the only course open for Spanish-English field testing, it was agreed. DLI undertook to obtain the requisite information from the Spanish embassy in Washington. At this writing, that information is not yet in hand, so our official contact with the Spanish military has yet to be made. For this reason, no concrete arrangements for site selection or preparation have been made, and therefore many details of the actual training and testing procedures remain to be worked out. In the following discussion, the reader should be aware of the tentative nature of the plans being presented.

For the record, it should be noted that our plans for Year 3's field test, at DLI-Monterey for English-Mandarin Chinese, were accepted by DLI in their draft form. This report is not the appropriate vehicle for the complete presentation of those plans, since we expect them to change as our research into that language pair proceeds and as our experience with the DLI environment broadens. We shall, however, give some Mandarin-related software some mention in the relevant section.

3.1 THE TRAINING MILIEU

The location for the S-E field test will be a Spanish Army base, probably located in Madrid, at which an intensive six-week basic English language course is taught by DLI-trained teachers. The students, Spanish military personnel, are being prepared for further training in English and some military specialty in the continental United States (CONUS). The Madrid course is their first formal exposure to English (insofar as that is possible within the cultural and educational environment of Spain). The course is designed to produce rudimentary English skill, if possible, and presumably to select those students having sufficient aptitude in language learning to justify the investment of further training time in the CONUS for some military specialty whose content is presented in English.

Staggered groups of Ss enter the school for a course of six weeks' duration. They are graded throughout this time by standard DLI testing and evaluation instruments and procedures, using the basic English curriculum provided by DLI. At the end of the course, their final grade—pass/fail—determines whether they are (a) shipped to the CONUS for further English language training at Lackland AFB, or (b) returned to the start of an incoming group's basic course.

We shall draw two matched groups of Ss from this pool, for as many successive six-week courses as are possible during the time the system is installed. Recycled Ss will not be used. The experimental group will be given full exposure to the system. The control group will be tested identically to the experimentals, but will receive no additional special treatment. Both groups will undergo the usual instruction offered by the school, with interference with the normal training of the experimental group being held to the lowest possible level.

3.2 OVERALL TIME FRAME

Since the course is six weeks in length and since the Ss are unavailable for testing both before and after this training, having not arrived or having been shipped out or recycled, all measurements must be made within the basic six weeks' time. This rules out the possibility of testing retention in any way, as was done in our previous pilot work.

We will have no contact with the students during their first week of classes. This time lag serves several functions: (a) it allows them to get into the normal routine of their work; (b) it allows them to be exposed to certain English-language materials in their classes and language laboratories, providing them with some rudimentary practice in English speech, behavior which is tested as part of the overall design; and (c) it allows their performance in class to be formally evaluated in terms of their one-week grade, a datum that will be of use in subject selection procedures.

Measurements of S behavior, of kinds to be described below, will be taken at at least four points in time: pretraining (1 week), two mid-training (3 and 4.5 weeks), and post-training (6 weeks). The mid-training times are provided in order to make possible the temporal delineation of any differential learning rates between the treatment groups.

Subject selection will be done immediately following the first week of classes, according to the procedures summarized below. For the balance of the training period, two groups of Ss will be trained/tested, using a staggered schedule set up with the aid of the base administration. The objective of this

schedule will be to minimize interference with important class periods for normal instruction, and to spread interference with other class functions across such functions within a given student.

3.3 TESTING PROCEDURE

On the four testing days, for all Ss, the following procedure will be used. In their entirety, they are not applicable to the problem of proper selection of Ss; the appropriate subset for this function will be presented in the next section.

All prospective (for the first testing session) or current (for all later testing sessions) subjects will be administered a three-part test. Its components are:

1. The Test of Aural Perception in English for Latin-American students (TAPEFLAS), developed by Robert Lado at the English Language Institute, University of Michigan (1957). This is a paper-and-pencil instrument that will be implemented via a prerecorded tape of an English talker speaking groups of words for the students to discriminate. The purpose of its administration is to provide a quick, objective measure of discrimination ability. This is thought to be highly correlated with the ability to produce proper speech discriminations. The test will be administered to groups of Ss sitting together, facilitating its scheduling and scoring.

The remaining two sections of the test are administered with the aid of the API, since they involve recording of speech. Ss are brought to the API facility singly and are given appropriate verbal instruction by the assistant and by the CRT display. They then speak and record two different types of English material.

2. Connected speech. One of the objectives of the research design is to determine whether Ss trained with a phoneme-level display can generalize their training onto larger phrases of connected speech. Testing this notion is quite difficult if Ss are allowed free rein in their English production, especially if one considers that the prospective Ss have but one week of English work on which to draw. Therefore, Ss will be asked to read English material used in their first week of classes. The same sentences will be used at all testing times so that an obvious cue to test time (from the judges' standpoint) can be avoided. The English sentences will appear on the CRT before the S, and he will have no immediate auditory model upon which to pattern his utterance; he will have to base it on his recollection of his first week's work.

3. Short utterances similar to training stimuli. This material will be most closely inspected, for it is here that we expect the experimental Ss to demonstrate the greatest effect of training. This is the only opportunity to collect such utterances from the control Ss, of course, since they receive no API exposure. As a partial control on the greater experience of the experimentals in producing such utterances, we have provided two measures: changing the overall recording format for the testing sessions, making it different from normal API training; and including unfamiliar (untrained) materials for both experimental and control Ss to record.

Two basic types of stimuli are used during the course of the Spanish-English training: minimal pairs and short phrases. Minimal pairs (MP's) are used to train production of and discrimination between vowel sounds, and aspirated and unaspirated initial stops; short phrases are used to train the suprasegmental features

of intonation and stress. A fuller description of the purpose and nature of the training and testing stimuli is given in Sections 4.2.1 and 4.2.2. Suffice it to say that, during testing sessions for experimental and control Ss, the MP-type material is expanded into a triplet format, to avoid introducing an intonation cue to the utterances. For example, if we are interested in measuring the degree of vowel discrimination S makes in producing the MP /beet-bit/, we will ask him to say /beet-bit-beet/, and we will ignore the last word. The short-phrase materials are recorded by S with a significant difference from the materials produced under item 2 above. In the entire present section, where all Ss record short utterances similar to the experimentals' training stimuli, they are provided with a prerecorded teacher model to emulate, as well as with a CRT display of the words they are to record. The recording and display are presented following the pressing of a button by S indicating his readiness to proceed to the next utterance. After the presentation of the teacher model, a 5-sec. "countdown," similar to that used in our pilot experiments, ensues to mitigate the mimicry effect observed then. No time constraints are placed on S's utterance thereafter; he is given a "go" signal after the countdown, and he can speak whenever he wishes. If he wants to hear the teacher model again, he may do so, but must then wait through another countdown.

The format and content of the test day procedures are identical across all Ss and testing days. Both groups will therefore become more familiar with the procedures and stimuli as time passes. This is an unavoidable fact, made necessary because of the need for samples of Ss' production of stimuli over training. A later section will discuss the means of analysis of these data to get at the overall effectiveness of the system. Test procedures were outlined here because a part of them serves the additional function of aiding in subject selection.

3.4 SUBJECT SELECTION

This process is designed to be executable quickly, to interpose minimum delay between testing and training of selected experimental Ss. The subject pool is defined as the entire class of beginning students. Two matched groups are to be formed within that sample. Two data from each student are used to produce the selection criterion.

1. One-week class grade. Instructors in the regular course will, through normal DLI procedures, assign a performance grade to each S. On-site negotiations will determine whether any special criteria are to be applied in this case. Students will be ranked by class standing, the better Ss receiving the lower numerical output. Ties will be averaged: i.e., if four students are tied for second place, they will all receive a rank of 3.5.

2. Discrimination-test score. The entire class will, by arrangement with the school, take the TAPEFLAS as a group at the conclusion of the first week of classes. Our field representative will grade the tests, and will assign a numerical rank to each S, as above.

The two class rankings will have equal weight for each S. Scattergram selection will be used. Depending on the number of experimental Ss that can be regularly scheduled for training within the time constraints worked out by our negotiations with base officiats, we shall select a wide-range sample of experimental and control Ss. For example: If there are 24 Ss in the pool, and we can train a maximum of 8 experimentals, we should divide the composite ranking of the 24 Ss into 8 groups of three Ss. One of each of the 8 triplets would be randomly selected as the experimental S; one would be assigned to a control slot; and the third

would be a control alternate, to take the place of the matched control should he drop out. There is no protection against an experimental S's dropout, since the system will be fully scheduled from the outset. The control alternates are tested in the same manner as the true controls for all testing sessions, to ensure that adequate control data for each of the experimental Ss are on record.

Hopefully, there will be more Ss in the pool than are needed for a full complement of experimental Ss, controls, and alternates. If this is so, it will cut down on the interference with the normal training of some Ss, who will thenceforth have no further contact with the research.

Though an ideal design would base selection of Ss on actual performance criteria tied to the accents of the potential Ss, this is obviously impractical in a field experiment performed under severe time constraints. The present method is at least based on some objective measures of Ss' aptitude and performance; the measures employed have face validity as correlates of accent and ability.

Following group selection, the triplets of Ss will undergo the second and third parts of the test-day procedures outlined above, where they record selected utterances in the API setting. These recordings are sent to BBN-Cambridge to serve as a baseline for later comparisons with the same S's performance later in training. Each matched group of Ss will be tested as close together in time as possible, to keep in-class training exposure balanced.

When the pretest procedures are completed for all triplets, training will begin immediately for the experimental Ss. The rationale, implementation, and content of the training sessions are described in a later section.

3.5 DATA REDUCTION PROCEDURES

At the conclusion of the six-week training cycle, we shall have the following data. For each of N experimental Ss and 2N controls, we shall have four sets of numbers, each derived from the TAPEFLAS as administered on the four test days; four sets of recordings of the English pronunciations of the Ss; and a complete set of class performance grades. Data from the alternate controls will be discarded, and the balance of the analysis performed with equal-sized matched groups of Ss. Several treatments of these data are proposed, and all will be carried out if practicable.

A. Subject by treatment analysis of variance (ANOVA) of the final DLI grades. This will be done on the composite grade, and again for any other available grade that might be more focused on behaviors of interest. We do not hold out much hope for the demonstration of a significant treatment effect, since course performance is a composite of many variables.

B. Subject by treatment by level (test days) ANOVA of the TAPEFLAS scores. This will provide a quick, objective answer to the question of differential improvement in discrimination ability as a function of API treatment. Recall that the experimental and control Ss are to be selected and balanced with pretest TAPEFLAS scores counting for half the rankings.

C. The samples of connected speech from all Ss will be collected in a randomized-order master judgment tape that will be played for a panel of experienced judges. Their responses will be collected, unscrambled, and used as the data for a subject by treatment by level ANOVA of the accentedness of the utterance. The first analysis will combine ratings of all utterances made by a given S on a given test day, and the level parameter of the ANOVA will be test day, as above. Later analyses, if significance is obtained in the primary one, will deal with specific stimulus materials for specific comparisons across test days, subjects, and treatments.

D. The shorter utterances approximating the experimentals' training stimuli will be gathered for subjective analysis onto a series of judgment tapes. The short-phrase materials for intonation training will be rated singly, and analyzed as above. The MP-type materials will be rated after truncation of the third word duplicating the first member of the MP. The resulting pairs will be rated, with higher scores going to those versions with better renditions of the intended discrepancy. From that point, the analysis will proceed along the lines used for the connected speech samples.

If time permits, we might be able to use our home system API for a limited evaluation of the short-utterance materials. The judgment tapes may be played back into the API, which may be configured to accept such prerecorded input and to subject it to the same types of analysis as would have been in force had the speech been produced in a normal teaching session—even though it might be the speech of a control S. The computer-based data plot might then be analyzed, either subjectively or objectively, and the set of such ratings might be analyzed statistically.

Specific attention will be paid, during all the above activities, to the following issues: (a) Are experimentals better after the same amount of training than the no-treatment controls? (b) Do experimentals improve faster than the controls (i.e., do they reach the point of diminishing returns earlier than the controls)? (c) What types of displays are the most successful, and what are the distinguishing features of a good display?

We have the capability of preparing new training materials and software with short delay, using our backup system at BBN-Cambridge, should that prove advisable when S or teacher reactions and suggestions are reported. The system as shipped need therefore not be the only version field-tested. That is the central virtue of a computer-controlled instructional facility: If the computer's peripherals are general enough to encompass the physical activities necessary, then the rest is manipulable through software change alone. It may thus be possible to iterate and improve the display techniques during the course of a series of field tests, to take maximum advantage of any information gained during the early stages of evaluation.

4. INSTRUCTIONAL SOFTWARE

4.1 THE HARDWARE

The configuration of the apparatus envisioned in Fig. 7, p. 19, BBN Report No. 2189, has been successfully realized for two systems: a backup home system to remain in Cambridge, and a field system for the coming tests. Figure 1 is a view of the equipment rack of the home system, and Figs. 2a and 2b show a student and the facilities within the sound-treated student room. Note especially the accelerometer mounted on S's throat, the head-mounted microphone, and the button box recessed within the work table. There exist only minor differences between backup and field systems.

Each of the buttons in the button box is internally lightable under program control, and we have adopted the convention that only "legal" buttons (i.e., buttons the program will respond to) are lit at any one time. Four colors of buttons are used, arranged in such a way as to minimize color similarity among neighboring buttons. During the course of utilization of the system, in training Ss and in system testing, several different programs are used. Each of these assigns different functions to the buttons. Ambiguities of function are eliminated by the use of overlay masks for the button box. Each program requiring a unique configuration of button assignments has its own overlay. Functions are indicated by short labels. A sample of such an overlay is shown below, where its specific functions are detailed (Fig. 4, p. 30).

As indicated in previous reports, the throat-mounted accelerometer is the first link in the newly developed pitch detector

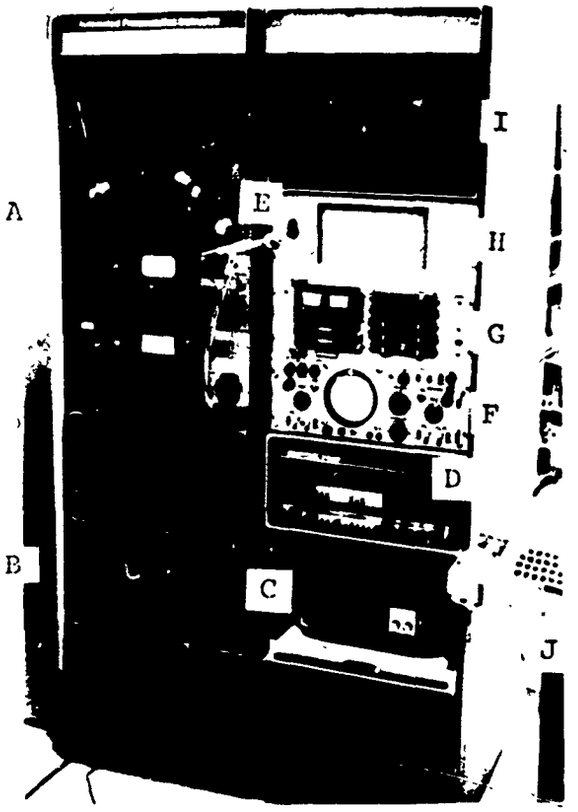

Fig. 1. Automated Pronunciation Instructor, Mark II: Equipment Rack

A. History tape recorder
B. Programmable interface, filter bank, misc. audio equipment
C. Datavoice bulk storage device
D. PDP-8E computer
E. Monitor's microphone
F. Slave oscilloscope
G. MacKenzie tape loop machine
H. Monitor's speaker
I. Multiplexer and A/D converter
J. Teletype

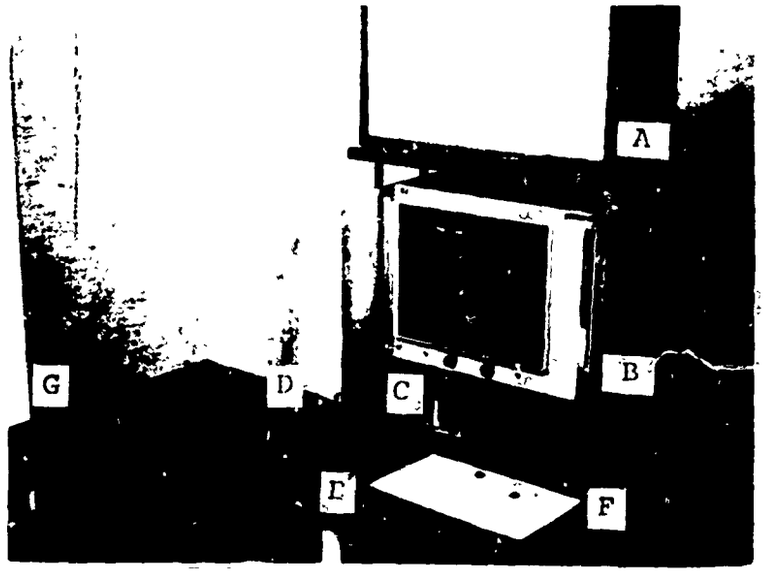

Fig. 2a. Automated Pronunciation Instructor, Mark II: Student Station

A. High-fidelity speaker
B. Display oscilloscope
C. Microphone power supply
D. Head-mountable microphone
E. Accelerometer for voice pitch detection
F. Button box with overlay
G. Work table

Fig. 2b. Detail of a Student Showing Proper Locations of Microphone and Throat-mounted Accelerometer

system. Information arising from this point is both analyzed in real time by the computer and recorded on two analog tape machines for later reanalysis, according to procedures outlined previously.

We have settled on the following parameters for the use of the Mackenzie Laboratories tape loop machine (mentioned in the previous report on pp. 24-25). The software to be described below can analyze utterances up to 2 sec. long. To allow for some freedom of utterance for S, and yet to minimize wasted time during loop playback, we use segments of tape 2-1/2 sec. in length. The tape loop, therefore, has two such segments—one reserved for S's efforts, and one for prerecorded teacher utterances. Each of these segments—and, in fact, all recordings used in the system—has two tracks: one track of standard audio information from the voice microphone, and another of pitch information derived from the pitch detector circuitry. The tray housing the tape loop possesses both record and playback capabilities. The teacher magazine is placed in another receptacle having only playback capability, to safeguard its contents. This repository of pronunciation models for both pitch contours and articulatory information could contain up to 7 minutes of recorded teacher speech; but considerations of storage space for the associated character display have limited the size of the teacher magazine to 24 2-1/2-sec. utterances. Each of these utterances, be it a minimal pair, single word, or short phrase, serves as the model for a subsection of S's work in a given session. Each is dubbed onto the tape loop, under software control, by S's command. He then works with that model, by methods specified below, until satisfied that the aural and visual comparisons between his and the teacher's productions are sufficiently close.

4.2 GENERAL SOFTWARE APPROACH

We have devised a general solution to the problems of displaying a variety of speech analyses for both student and teacher, in real time and in an internally consistent manner. To expand on those goals: whereas in previous research we had no recourse to the acoustic analysis of teacher models to provide targets for S's speech, the new tape systems gave that opportunity. With that opportunity came the inevitable problem of how to display such information. Another problem concerns the training of unsophisticated Ss in the minimization of more than one type of accent, since each new type requires a different display algorithm. In order to minimize negative transfer from one display to the next, commonality of system operation procedures should be maximized; that is, a given button should do the same thing across displays. Finally, a way was needed to make all of these constraints operative within the context of time-plotted information presented on the display, since previous research had shown that this display mode was most easily understood by naive Ss.

Our approach is based on the premise that many useful displays can be produced by plotting the output of certain speech analysis algorithms as functions of time. All the software for the display of several different analysis algorithms has been made mutually intercompatible. This generality arises from common data acquisition and display sections, with the differing algorithms being called for in a flexible way in real time.

Two seconds of speech from either the student (S) or teacher model (T) exist within the program as 200 spectral frames. Each such frame contains the log energy derived from the 19 bandpass

filters, giving a 19-point spectrum of the speech signal for that sample period, and a single additional number specifying the most current pitch period, if any. This latter number can range from the equivalent of 50 Hz through 800 Hz, the limits of action of the pitch detector, on a logarithmic scale. That is, equal amounts of change in the pitch quantity, as stored, signal equal ratios of pitch change.

Figure 3 illustrates this information for four such speech samples as recorded from a male speaker uttering "beet-bet" with a rising inflection on the first and a falling inflection on the second word. The four frames were selected to fall within the voiced segments of each word, one each at the start and finish of each vowel. Note first the change in recorded pitch during each vowel, and second the relative independence of the spectral envelope from the exciting frequency of the glottal signal. The differences between /i/ and /ɛ/ are far more extensive than the change in either wrought by shifting pitch. This is a demonstration of the utter impossibility of deriving pitch from the output of the present filter bank, and of the utility of the approach used.

As is easily imagined, the amount of space necessary for the storage of 200 such frames is considerable. Provision for full storage for the last 2 sec. of both S's and T's speech is thus not made. Rather, the program stores that amount of information about speech of the last speaker, be it S or T, and performs all analyses upon it before moving on to the next operation. Since the actual spectral frames, with all their detailed and confusing information, are never displayed to S, this loss of the spectral information concerning T's model utterance upon S's next input is immaterial. The program has already stored away the output of the crucial algorithms for comparison with S's corresponding algorithms. (In

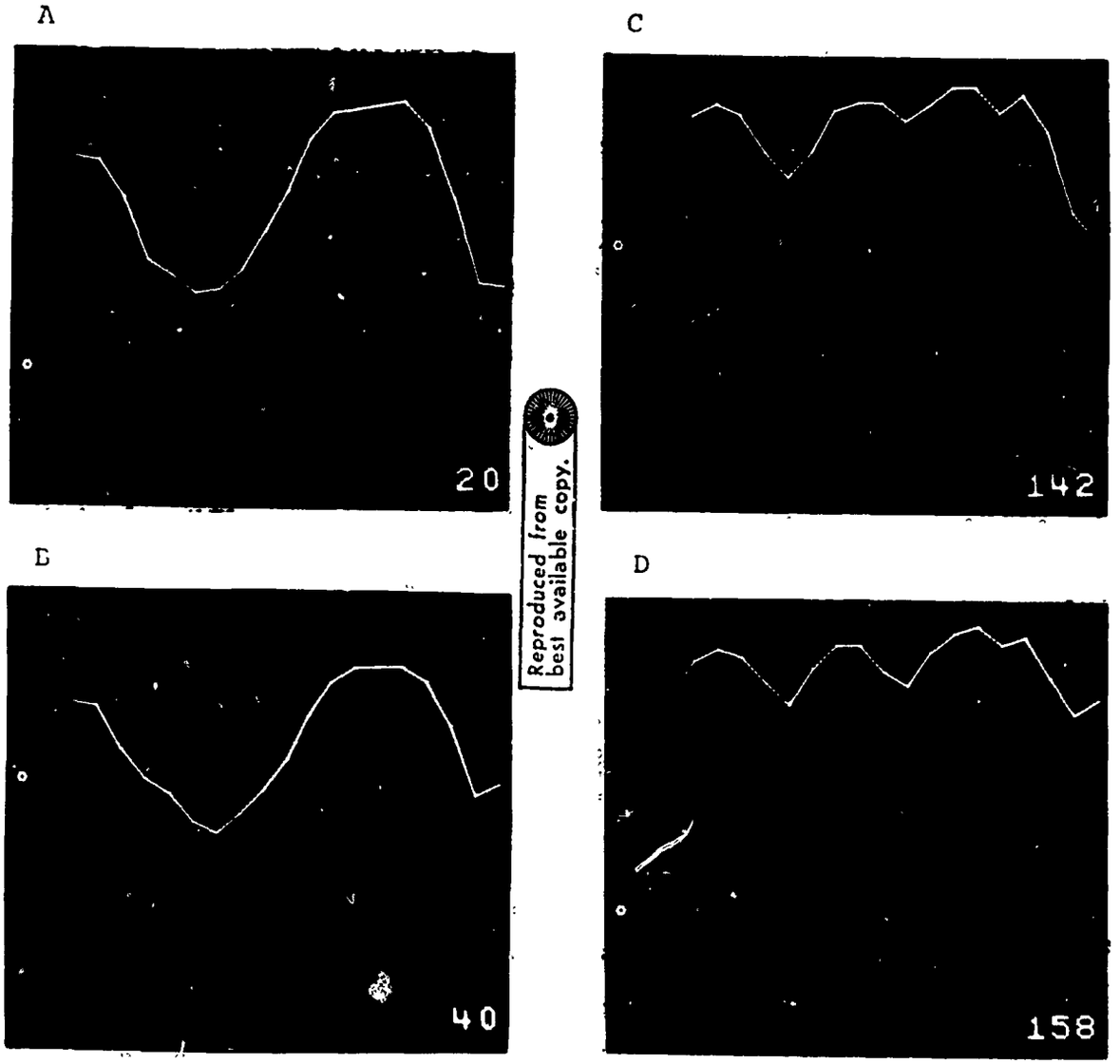

Fig. 3. Four Spectral Frames from the Utterance "beet-bet." (See text for details.)

Key to interpretation of the plotted data:
The abscissa is filter number; there are 19 points.
The closely spaced ordinate dots are proportional to a dB scale of energy as read from each of the filters.
The curve is plotted with respect to the above; there are 19 data points. Formants are visible as "bumps" in the spectra.
The widely spaced ordinate dots represent the log scale of pitch. Reading up, they correspond to octaves of 50,100,200, & 400 Hz.
The pitch recorded for the particular spectral frame is indicated by the position of the small circle. E.G., Fig. 3a's pitch is slightly above 100 Hz.
The number at the lower left gives the number of the spectral frame. Since the sampling rate is 100/sec., each frame covers 10 msec.

fact, the actual spectral frames themselves need not be stored at all, since all algorithms can be computed and stored in real time; but the point is moot, since we desired 2 sec.'s worth of stored frames for systems work. Later versions may eliminate this feature if needed.)

The program associates three crucial algorithms with each speaker at any one time. They are: (a) Loudness—a weighted sum of the energies across the spectrum, computed in such a way as to respond proportionately to vocal effort, independent of the speech sound. (b) Pitch—being the output of the pitch detector after conversion to a log scale of frequency. If, for a given sample, there is no voicing information, this datum is zero. The pitch information reflects the last completed pitch period resulting in digital output at the final stage of the pitch detection logic and sampled by the program. There are 100 of these samples made per second, so we do not recover a period-by-period record of the pitch excursions of the voice. We obtain pitch information with the same resolution as spectral information. The grain is sufficiently small. (c) A variable assignment, depending on the pronunciation problem that is the subject of the current stimulus. There are several candidates for this position; it may be filled with a tongue-position function or with an aspiration function. The logic for the assignment and for the use of the other two functions will be explained next.

The loudness function is used in every display, though it is almost never directly displayed. Its chief purpose is the signaling that an utterance has begun, whether it be speech starting with a voiced or unvoiced sound. When the system is ready to accept input, the start of speech is usually marked at the point in time where the loudness function first exceeded a threshold.

The pitch function is used only for those displays requiring attention to pitch contours, such as suprasegmental intonation patterns in English or for segmental tone drill in Mandarin Chinese.

The third function is used whenever the two above functions are irrelevant to the parameter being trained. Its use is best understood following a consideration of all displays currently being developed in both the Spanish-English and the English-Mandarin language pairs. Complete descriptions of the curricula and full listings of the training stimuli are unavailable at this writing, but the outlines and strategies are firm. Examples of stimuli appropriate for each type will be given.

4.2.1 Spanish-English Displays

The S-E curriculum addresses itself to three distinct accent parameters: vowels (tense-lax, diphthongized, and diphthongs) in monosyllabic words, aspiration of initial consonants, and prosodic intonation and emphasis. The first two parameters (vowels and aspiration) are taught via a minimal pair format.

4.2.1.1 Vowels, tense/lax. Here we contrast monosyllabic words containing one of the five tense Spanish vowels with minimally differing English words. Examples: beet-bit, boat-but. Algorithms designed for these distinctions are of the tongue-position sort. They may be general first-formant functions, indicating tongue height; they may be general second-formant functions, indicating tongue front-back; or they may be special-purpose functions designed to maximize the particular distinction being trained.

4.2.1.2 Vowels, low/mid. The vowels /æ/ and /ɔ/ are contrasted with their nearest Spanish neighbors. Examples: hot-hat, base-bass, so-sir, hate-hurt. Display algorithms are selected from the above-mentioned set.

4.2.1.3 Vowels, diphthongs/dipthongized. The emphasis in the above stimuli is to get S to move his tongue to new and different (for him) regions of the mouth, but not to produce too much tongue movement during the voicing. Such movements produce diphthongs, speech sounds more numerous in English than in Spanish. The explicit display of a time axis in the Mark II API makes it possible, for the first time, to incorporate these new phonemes into a training procedure. Tongue gestures produce curved traces in time, traces that move from one tongue position to another during the diphthong. Here, we train S to recognize the fact that English /e/ and /o/ are diphthongized: beet-bait, shoe-show. We also train his production of the diphthongs /ai/, /au/, and /ɔi/: heed-hide, sod-side, boot-bout, shot-shout, so-soy, see-soy. Note that whenever possible, the above stimuli are presented with minimally differing stimuli at both ends of the tongue gesture: i.e., the diphthong /au/ is presented in the context of /a/ and /u/ pairs. Display algorithms are selected as above, in ways to highlight the gestures required.

4.2.1.4 Aspiration of initial consonants. As noted in previous reports, the phonemes /p/, /t/, and /k/ are difficult for Spanish speakers whose versions often verge on /b/, /d/, and /g/ for English listeners. The accent arises due to the speaker's producing too short a time interval between the onset of the word and the onset of voicing, and in not producing sufficient aspiration noise during that voiceless interval. The aspiration-detection algorithm used previously is employed here in a minimal-pair context: deem-teem, beak-peek, goat-coat. The amount of aspiration beginning each utterance is shown for both S and T, as well as the onset and duration of voicing. S's job is the matching of his pattern with that obtained from the teacher's prerecorded utterance.

Note that in all the minimal-pair work described above, one member of the pair is always a word containing a phoneme familiar to the Spanish speaker. His task is to work from that base to a new phoneme, using his familiar word as a point of comparison. This will be made clearer when the MATCH function is described.

We plan a progressive increase in the difficulty of the curriculum, within the limits of time. Upon their first exposure to the above types of stimuli, the material is grouped, with, e.g., all occurrences of the /i/-/I/ distinction in one section of the training stimuli. Further, the distinctions made in the pairs always move from Spanish to English. The word containing the familiar phoneme always precedes the more difficult member of the pair. Later exposures of Ss to the material will involve scrambled orders of stimuli. Other parameters, such as display algorithm used, will remain unchanged.

4.2.1.5 Intonation. We will train Ss in the proper production of short-phrase and short-sentence material using the pitch detector. It has been observed that Spanish-speaking Ss pronounce English words, phrases, and sentences with Spanish intonation and emphasis, and that their suprasegmental pitch contours sometimes fail to carry the desired meanings. The present display techniques may help draw S's attention to the differences, and, by comparing his pitch (and therefore emphasis) contours with those of the T model, his intonation may be improved. We have selected materials from the word level (airplane) through the short-phrase level (railroad station), to declarative (Good morning, I bought a suit), and interrogative sentences (How are You? May I help you?). A simple display of pitch as a function of time will suffice for most of this work although we are still considering a mode where the loudness of the speech is plotted as a symmetrical envelope around the pitch trace. This is, however, both partially redundant and visually confusing, so its use will be limited if adopted.

4.2.2 Displays for English-Mandarin Chinese

We expect that the fundamental problem confronting the English speaker in this language pair will be pitch control. We have described above a general software frame within which pitch contours for the speech of S and T may be derived and plotted in real time, so the basic work for that speech parameter has been done in a manner compatible with all API display software. All that is required to bring that phase of an English-Mandarin Chinese curriculum to readiness is an appropriately graded curriculum. A draft outline of the pitch curriculum will be described now. Displays for unfamiliar vowels and consonant clusters in Mandarin remain to be developed. It is anticipated that they will be cast into the time-plotting mode.

The first level of pitch display is monosyllabic words, minimal-pair mode. All the permutations and combinations of the four Mandarin tones are presented for S's mimicry. Simple carrier syllables are used. Examples: lā-lá, nǐ-nì. The exercise is highly similar to one on p. 3 of Hockett's "Progressive exercises in Chinese pronunciation." S learns the proper interrelationships between the tone levels by intercomparing his and T's versions of the MP's with the match function described below. Since the loudness pattern of some of the Mandarin tones is crucial, some training at the basic level may be administered with the loudness function plotted symmetrically around the pitch. It may well prove useful when the stimuli are sufficiently simple.

Following the drill on tones in isolation, Ss will move to the API-assisted drill on the structure of short phrases. The MP mode is abandoned as S mimics T's prerecorded versions of phrases such as: Jūnggwo bīng, Měigwo màudz. Combinations of tones, phrase

rhythm, and the pronunciation of the zero final are trained at this level. The examples given are from DLI's Basic Course in Chinese-Mandarin, Lesson 3.

A further phase of tone training involves the production of short sentences, such as: Jèi shr̀ shū. Nèi bú̀shr bàu. The complexity of the material could be increased and interrogative particles added as desired, within the limitations of a 2-sec. utterance. This limit is probably close to the capacity of S to read a long pitch contour, and most contours of interest can be drilled within that time span. The above examples were also taken from DLI's basic course.

Before the final decisions on stimulus materials are made, we will undertake further research on the speech of Mandarin informants available to us here in Cambridge. Also, we will consult with trained teachers of Mandarin, both at neighboring schools and at DLI, to select appropriate accent parameters for further display development. Whenever possible, training materials will be taken from the standard DLI curriculum. That will ease the methodological problems of interpreting the data obtained from experimental and control Ss.

4.3 SESSION SCENARIO

4.3.1 Introduction

The present section describes the options available to, and the typical actions taken by, a subject using the system to work on the reduction of his accent in any of the above-mentioned parameters. The presentation will be complicated by the need for the reader to keep in mind that the same software framework is used for the

display of widely disparate speech analysis functions, and at times different actions are taken by the program, depending on which parameter is in use. Two representative stimulus types are chosen to illustrate the general capabilities of the software, and the displays for those parameters are described and pictured (where possible) in parallel as the presentation moves through the various operations.

As a representative of the plotting of algorithms arising from the filter bank, the tongue-height function (sum of filters 1 and 2 minus sum of filters 3 and 4) is chosen. It is presented within the context of a pair of words drawn from the diphthong minimal-pair training materials (Section 4.2.1.3 above): "heed-hide." This serves also as the representative of all minimal-pair stimuli, since they are treated alike in display manipulation.

The pitch detector provides the second speech-analysis function displayed in the scenario. Since we have already chosen a minimal-pair type stimulus for the first example, a phrase-mode sample is used next. It could have been drawn from the suprasegmental intonation drills in the Spanish-English language pair, but a Mandarin sentence of the type mentioned in Section 4.2.2 above will be used: Jèi bú shr̀ jūng. It means: "This isn't a clock."

In both cases, informants fluent in the production of the model sentences are used, and Ss with the appropriate language background attempt to mimic the Ts' models. Specifically: a male native American is the prerecorded teacher whose version of "heed-hide" is the model for a male native Spanish speaker; a male native Mandarin speaker models "Jèi bú shr̀ jūng" for a male native American.

Of course, the examples below are isolated in the sense that they are part of a large number of similar utterances for the same accent problem, in the sense that there are other problems with rather dissimilar stimuli and in the sense that a static picture gives only a faint flavor of the interactivity of the display. It is hoped that the introduction given in Section 4.2 above will impart some picture of the extent of the possibilities. The flexibility of action available to S will be described, as possible, within a linear exposition of his options.

4.3.2 System Setup

The field technician prepares the system for the arrival of S by consulting the experimental protocol and determining what material is to be covered in the coming sessions. He loads the computer with the appropriate software from the digital bulk storage tape, and sets any needed run-time parameters using the computer's switch register. He loads a teacher tape magazine into the playback channel and sets it in such a way that the first training stimulus is in position for the start of work. He loads into the computer a separate subsection of information about the word list spoken by T, so that the software may display alphabetic characters on the CRT for S to read while he works. Finally, the technician mounts a history tape, moves it to a blank spot, and marks the start of the session's tape via an audio cue giving S's name, the date, the stimulus list, and any special parameters. A record of every utterance placed on the tape loop will thus be available for later perusal by the analysis team at BBN-Cambridge.

When S arrives, the technician puts on the microphone and attaches the accelerometer, if pitch information is required by the software to be run that day. If S requires any instruction

over that given him at the first session, the technician can give it at setup time or at any time during the session via the HELP button.

The button box has a standard overlay used on all training days with the general software package. Seven of the 12 buttons are functional in standard mode, with the other five "privileged" to systems personnel only, for special-purpose diagnostic functions. Figure 4 illustrates this overlay in the version used by English-speaking Ss. Spanish-language button labels would be used where indicated. What follows is an exposition of the functions of each of the buttons, presented in a hierarchical-sequential order. At some point in the description of each button, the various options open to S will be listed, since not every function is available to S at every juncture.

4.3.3 NEXT Button

This is the only illuminated button at the start of the program. It is pressed at the beginning and, later, whenever permitted by the software and desired by S. Its function is to cease work with any previous stimulus and to proceed to the next model utterance. The system dubs the next 2-1/2 sec. segment of the T magazine onto one segment of the tape loop, while simultaneously listening to the output of the T magazine with the filter bank and pitch detector. Algorithms appropriate for the software are computed and stored over a 2-sec. range for the T speech. During this period, S hears nothing. The history tape recorder is activated and records the T speech as well, to mark the tape when S begins work on the next stimulus. S's CRT display is changed when NEXT is pressed; any preexisting functions are removed, and the new stimulus appears at the top left for S's immediate inspection.

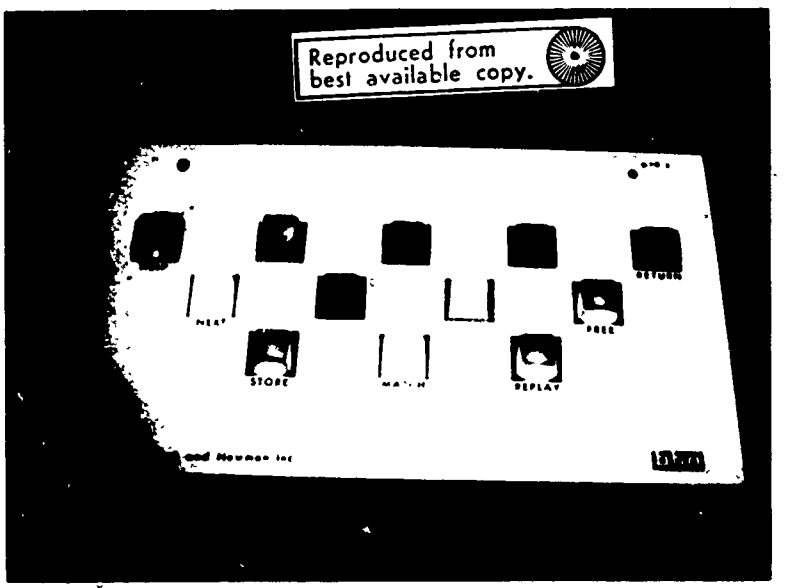

Fig. 4. Student Button Box with Standard English Overlay.

At any point during program operation, legal functions are indicated by lighting of the button. The illustrated pattern is shown, among other times, following operation of the STORE function. The five central buttons are "privileged" (see text, Section 4.3.9). Privileged-function overlays are also available.

Figures 5a and 5b show the display as it appears at this point for both the examples.

S is not allowed to hear T's version of the stimulus or see the time plot of the computer's analysis of that speech, so that his first attempt at each stimulus in each new session might be unbiased by the mimicry effect. Thus, the history tapes will contain useful data on intra-session improvements.

At the conclusion of the 2-1/2 sec. of tape dubbing, the "ready" sign (the small cross) at the lower right of the display) reappears (it is absent whenever the program is performing a function that is noninterruptable), and a subset of the button-box lights is illuminated, telling S that he may proceed to new activity. Legal activities following NEXT are STORE and HELP.

4.3.4 STORE Button

S presses this when he wants an opportunity to record his attempt at the stimulus on the tape loop, and to view a real-time analysis of his speech. The following things happen. The tape loop, which in its resting state is positioned just before S's segment of its two 2-1/2 sec. parts, is set in motion and its record functions are enabled and hooked up to S's microphone and accelerometer via the audio circuitry and computer-controlled switches. The history tape is similarly activated. The button-box lights go out and the + sign is removed from the display. S has 2-1/2 sec. in which to speak the stimulus. For the first 1/2 sec., the software awaits a suprathreshold speech sound. If it hears something, the 2 sec. of speech begins at that sample. If nothing is said within the first 1/2 sec., the 200 samples are filled anyway and whatever speech is encountered will be analyzed. This arrangement

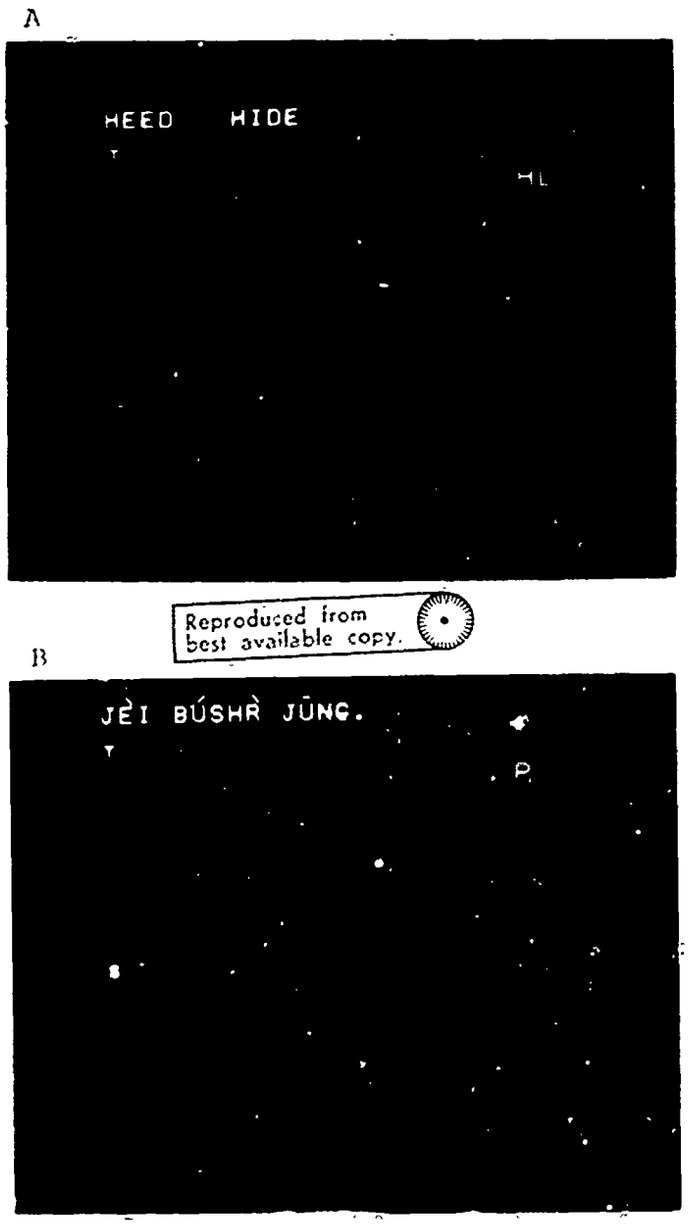

Fig. 5. CRT Display following [illegible] Pattern for Two Types of Stimuli. (See text for explanation.)

gives increased temporal flexibility while retaining the maximum allowable number of 200 spectral samples.

In real time, the program reads the preprocessors, stores the data, and plots a time display of the parameter of interest on the lower portion of the screen (under the mnemonic abbreviation S). Figures 6a and 6b show the display as it appears at the conclusion of S's speech. The tongue-height function plotted in Fig. 6a shows the discrepancy in location and gesture between the two utterances of S; the pitch function in Fig. 6b represents the time course of "Adam's apple position" during the production of the Mandarin phrase.

4.3.4.1 Special treatment of data. Now that actual time plots of algorithm outputs have been shown, two additional points regarding the display algorithms will be made. They involve means we have found appropriate to improve the legibility of the display.

We have observed that all time functions, be they pitch- or filter-bank-derived, have a certain variance in addition to overall trends. While pitch may be increasing over a period of 50 samples—1/2 sec.—its increase may be apparently grainy, due to natural variability in glottal action and to interactions between sampling time and wave phase. Functions arising from the filter bank are similarly grainy, since the contents of any filter may show time variations in phase with the excitation of the glottal pulses; hence, any function based on sums and differences of filter energies within a given spectral frame will show time variance about a central value or trend. To combat this, we have applied a digital smoothing algorithm to all functions of the general form:

$$f_n^* = \frac{Af_{n-1}^* + Bf_n}{A + B} \qquad (1)$$

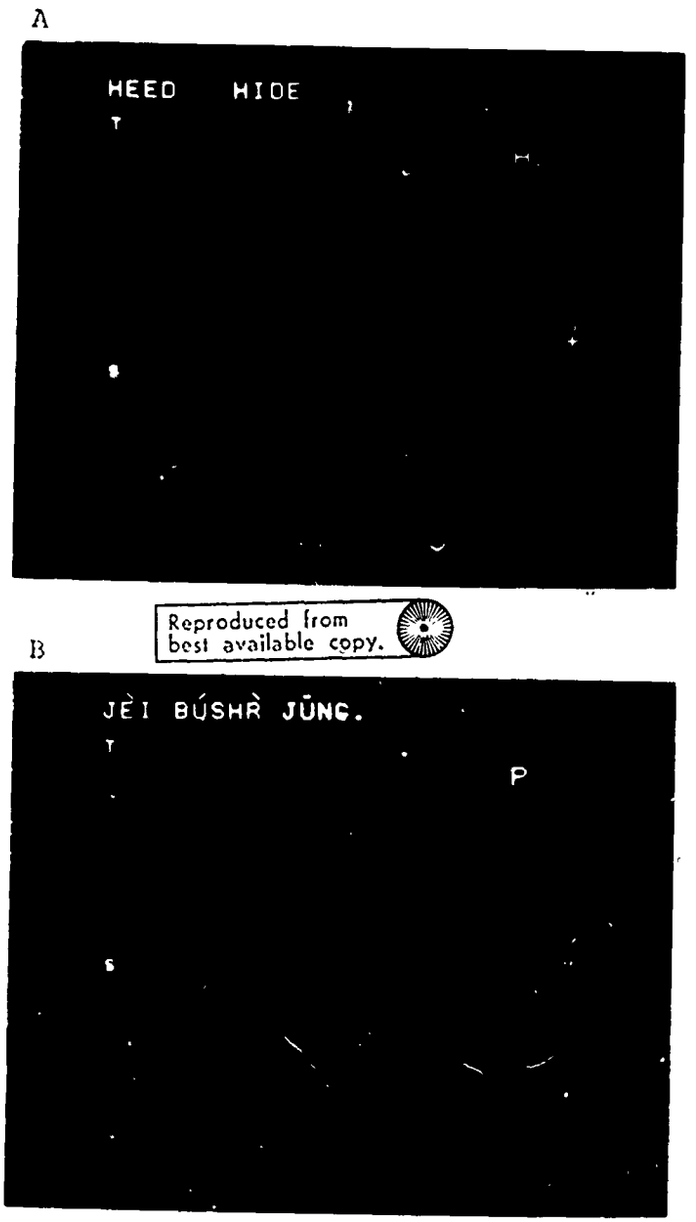

Fig. 6. CRT Display following Student Speech in STOP Function for Two Types of Stimuli. (See text for explanation.)

where f_n^* is the output function at sample time n, and f_n is the input value derived from the pitch or tongue-position function. This is thus a recursive function where the constants A and B determine how much the previous sample's output weighs in the determination of the present sample's output value. The constants can be chosen to cover a wide range of smoothing time-constants; those in use here are in the 15-msec. range. This digital smoothing is far superior to that which might be provided by merely adding more analog filtering to the input channels, for two reasons: (1) it is infinitely variable by software change alone, and (2) it is applicable to computed function output rather than to raw input data. The latter advantage is clearly necessary in light of the consideration that both pitch and tongue-position functions have sharp time onsets and offsets. If pure filtration were used, such transitions would produce gradual smoothed (negatively accelerated, exponential rise and fall) curves, when, in reality, the speech was either unambiguously present or absent. Two additional constraints are therefore placed on Eq. (1):

$$f_n^* = 0 \text{ if } f_n = 0 \qquad (2)$$

Thus, if any frame possesses a function value of zero, the smoothed output is set equal to zero. The first zero frame following a period where there had been algorithm output results in immediate cessation of function plotting.

$$f_n^* = f_n \text{ when } f_{n-1} = 0 \qquad (3)$$

Thus, the first frame at which the algorithm becomes nonzero produces a full-valued smoothed output, with no recursion, with previous zero values.

Digital smoothing is not the complete answer to the needs for improved display legibility. The second stratagem we have employed might best be called a "garbage collector." Operating on the 200 smoothed function points, produced as described above, the GC algorithm proceeds to inspect the data. If it finds a group of four or less nonzero points delimited by zero values, it removes it outright. For every other group of five or more contiguous points, it removes the first two. There remain no groups smaller than three contiguous samples (an unlikely result), and those that remain have had 20 msec. of data removed from their leading edges. The reasons for the GC are twofold.

The first reason is background and mouth noise. Small non-speech sounds are often picked up by the preprocessors. They may be too small to trigger any thresholds that actually start data storage, but, once such storage is underway, the software has few ways of discriminating noise from speech. This may result in function output during "silent" periods. Such noise output is usually of short duration, however: easy game for the GC.

The second reason is start-up transients. When voicing begins, the first one or two function outputs are often different from those coming later. This is particularly true of the pitch data, due to physical phenomena occurring at the start of vocal-cord vibration. We deemed the loss of two samples' data a reasonable price to pay, if the gain was increased display legibility. One-fiftieth second is imperceptible in all real-time playback contexts.

Let us return to a description of the further activities of the system in the STORE function. The reader will recall that we left the account at the stage where S has finished his utterance, and it has been: (a) stored in analog form on the tape loop, with

a pitch and a voice track, (b) similarly stored on the history tape, (c) digitally stored within the CPU, and (d) displayed in real time in the relevant section of the CRT.

Now is the time for S to hear T's version of the utterance. If this is his first press of STORE for this word, he will hear and see T for the first time. The tape loop is started once again, this time in playback mode, and S hears the model utterance through the speaker mounted over the CRT. At the same time as the speech is heard, S is gradually shown the time-trace of the T utterance that the program computed when the NEXT button was pressed. Figures 7a and 7b illustrate the display at the conclusion of this cycle. The entire STORE sequence takes 5 sec. When S speaks during the first 2-1/2 sec., he produces a time plot as he speaks, giving himself immediate articulatory feedback. When S listens to T during the second portion, T's time trace appears, in real time, in synchrony with T's speech. It is as if T were sitting with S, plotting his data in the same manner as does S; but in reality this is implemented by a procedure that "discloses" successive 10-msec. samples of T speech as the T segment is being played.

Figures 7a and 7b, then, are a visual record of the audio contents of the tape loop following the STORE function. S then may inspect his and T's plots for similarities and differences, and plan his next actions accordingly. The button box lights and the + sign reappear. Legal activities following STORE are NEXT, HELP, STORE, REPLAY, MATCH, and FREE. S may go on to the NEXT stimulus if satisfied with his performance. He may desire to repeat STORE. In this case, the T trace remains on the screen during the time S may record his attempts. Now that we have recorded an unbiased attempt by S to produce the utterance, he might as well have the benefit of a visual model to shoot for during his attempt.

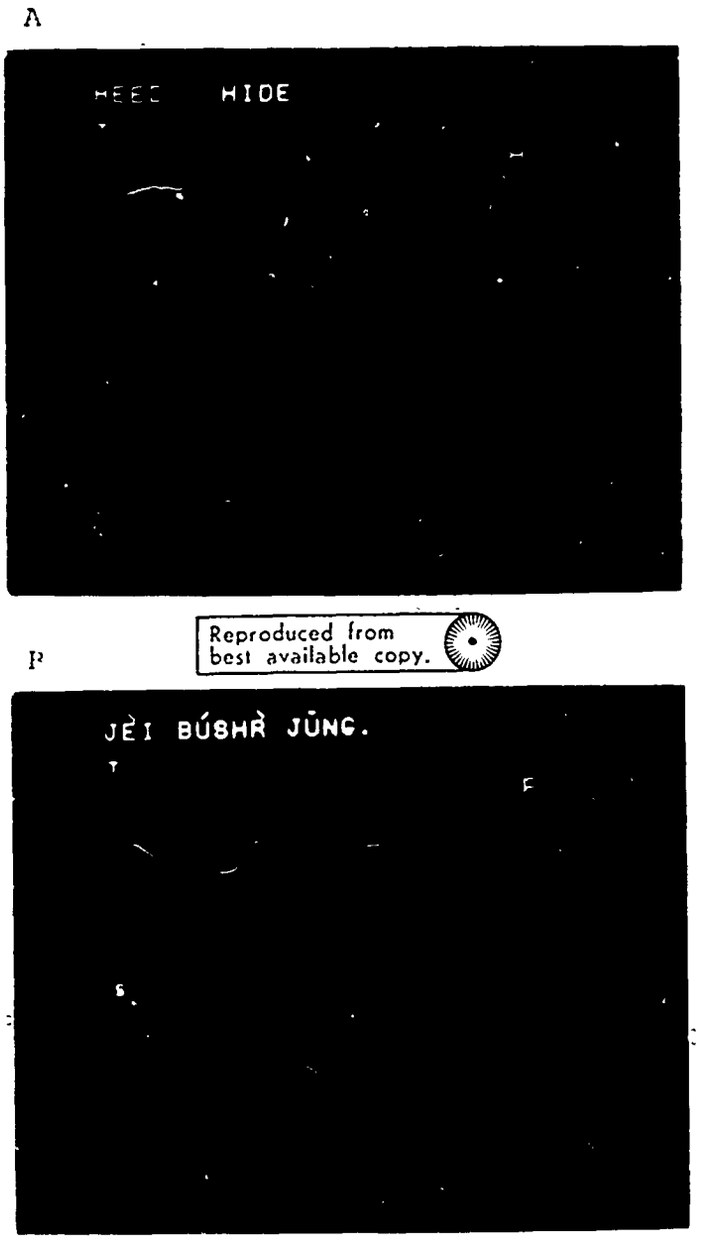

Fig. 7. CRT Display following Complete Operation of the STORE Function for Two Types of Stimuli. (See text for explanation.)

4.3.5 REPLAY Button

The tape loop now contains S's and T's versions of the stimulus, and the CRT shows a static display of the relevant parameter as a function of time. S may elect to cycle the tape loop to make a comparison between the two utterances. This is simply effected via the actuation of the playback circuitry of the loop. First, the last S utterance recorded in STORE is heard, since the loop is left at the start of S's segment following the STORE sequence. Next, T's prerecorded utterance is heard.

A special feature of the REPLAY function is real-time redisplay of the data. Whereas it would have been simpler to just leave the display static, in the form it took at the conclusion of STORE (Figs. 7a and 7b), a dynamic aspect is provided. At the start of REPLAY, the S trace is deleted and then replaced in real time as S's speech is played back: the display is replotted, just as it was when S spoke originally, but without S's being required to speak. Then, when S's section of the display is complete, T's trace is removed and replaced as above. This gives S the opportunity to reassociate specific parts of his and T's utterances with specific sounds and portions of the display. He can pinpoint just what it sounds like when a particular feature of the display is drawn.

Legal activities following REPLAY are NEXT, HELP, STORE, REPLAY, MATCH, and FREE.

4.3.6 FREE Button

One of the minor inconveniences of the STORE function is that, after S presses it, he has a maximum of 2-1/2 sec. in which to speak, and only 2 sec. of that interval will be displayed.

This is a considerably longer time than the 1-sec. total and 1/2-sec. analysis provided in the Mark I API, but we expect that it will, nevertheless, engender some detectable anxiety in naive Ss. Of course, time limitations are unavoidable when a tape loop is used in the present configuration, and some limits are essential to efficient system operation. There should, however, be some means to bypass the analog recording process and allow S free access to the digital analysis facilities of the system.

Pressing FREE connects the microphone and pitch detector to the preprocessors and replaces the (+) symbol with an (*), indicating system readiness to accept new S data. Time plots already on the screen remain unchanged. The flicker of the display increases, since the program is both maintaining the display and "listening" to the filter bank and pitch detector in an attempt to determine when S begins talking. One hundred spectral frames per second are inspected.

When the loudness function exceeds a given threshold or when the pitch detector outputs a nonzero value, that frame is taken as the start of speech. The S display that had been present is erased immediately and a new one constructed. A small cursor below the baseline moves across the screen, covering the full distance in 2 sec. Any sample producing function output is plotted in real time over the cursor's current position. The (*) is present throughout this process as a further indication of the readiness of the display to listen to S's speech. Note that nothing happens to the display <u>until</u> S begins to speak. He may "gather his wits" before speaking, and begin at his leisure. Once he begins, the system, which has been listening all along, stores 2 sec. of speech and then halts for a "refractory period."

When the cursor reaches the rightmost extremity, the (*) disappears for 1/2 sec., during which time the display remains unchanged and unalterable. This "refractory period" is to allow S to stop speaking; if it were possible to recycle the cursor to the left immediately the S filled the 200th bin, then he might erase what he had plotted by running over the allotted 2 sec. When the (*) reappears, the system is in the initial state, and ready to accept new speech and to erase the previous S trace. The contents of the T trace are not touched in any way during FREE; only the NEXT button can alter them. No illustrations of the display configuration are given, since it is identical to that shown following STORE. The dynamic, unconstrained, real-time nature of the FREE mode can only be described verbally. S speaks whenever he wishes, without having to preface each utterance with a press of a button.

Exit from the FREE mode may be done at any time. The only legal button, once S enters FREE, is RETURN. The contents of the S trace remain unchanged when S exits FREE: whatever they were at the instant RETURN was pressed. When this happens, there is a discrepancy between the contents of the tape loop and the time trace of S's last utterance in FREE. This is immaterial to all functions save REPLAY, and hence REPLAY is not enabled when S returns from FREE mode.

The legal functions following RETURN are: HELP, NEXT, STORE, MATCH, and FREE.

As just described, the system does not allow S to make an audio comparison of the contents of the two segments of the tape loop following use of FREE mode, since the sound and the picture will not agree. This is the price one pays for freedom

from tape-loop time constraints. Nevertheless, there still remains the possibility for purely visual comparisons between the two traces: the MATCH function.

4.3.7 MATCH Button

This is of central importance to the operation of the Mark II API, since it makes explicit the target that S is attempting to imitate, and considerably eases his pattern-recognition task. In our previous research, adequacy of S's rendition of a sound was indicated by the plotting of S's speech analysis with respect to predetermined nominal values (acceptable ranges of F_1-F_2 space, or acceptable voice-onset times and aspiration levels). Now, we are asking S to compare time plots of a relevant speech parameter generated by himself and by T, and to attempt to minimize the discrepancies. What are the various ways in which S's and T's traces may not be similar, and how do they arise?

(a) Different utterance onset times within the 2-sec. periods allotted each speaker. Once the loop starts, actual speech may begin anywhere. Furthermore, in minimal-pair or multi-word utterances, the inter-word wait time may vary between speakers.

(b) Different average function values. If a tongue-position function is being plotted, acceptable versions of a given vowel phoneme may produce function values that differ because of the physical sizes of the two vocal tracts. If pitch is being plotted, speaker pairs having different average fundamental frequencies will produce data at different distances from a zero baseline. In both cases, intra-speaker discrepancies will be consistent, but inter-speaker comparisons may be open to misinterpretation.

(c) Different utterance durations. Once begun, S's utterance(s) may produce more data points than comparable portions of T's trace.

(d) Different trace shapes. The time courses of the gestures may differ.

The latter pair of areas are where the immense pattern-recognition capabilities of S are best called into play, aided by appropriate verbal instructions on the nature of important discrepancies. The former pair of causes for S-T trace dissimilarity can cause unnecessary confusion to S on inspecting the standard display as illustrated in Figs. 7a and 7b; the MATCH function makes things straightforward by eliminating them.

4.3.7.1 Intra-speaker MATCH mode. By introducing the feature of analysis and display of T speech, we have compounded the problem of rationalizing different vocal tracts within the same display. However, this issue may be neatly avoided in a large number of cases due to the use of minimal-pair stimuli. Recall that in the Spanish-English vowel-training work, a ground rule is that each pair shall have one word with a Spanish home vowel and one new English-vowel word. This gives a fine baseline for comparison purposes. For minimal-pair pitch training in Mandarin, the general level of pitch is less critical than the gesture shapes; and the proper overall level is quickly set by S after minimal exposure to tone structure. In other words, if S can be depended upon to provide a reasonable attempt at either the overall level of a function or a reliable version of one member of a minimal pair, then the most important aspect of his display is not its parts' distance from a zero baseline, but the distinction between the two members of the pair in level and/or shape and/or duration.

For minimal-pair type stimuli, therefore, software is provided that produces the following effects when MATCH is pressed. The trace produced by the first member of a minimal pair remains untouched. The second member is shifted leftward in a smooth, negatively accelerated movement that places its first point in X-axis conjunction with the corresponding point of the first member. The transition itself takes about 1 sec.; it remains in overlaid position another 1 sec.; and it returns smoothly to its starting point over the same interval. Figure 8a illustrates the overlaid position. Note that the software has sensed the onsets of all four words being displayed and has matched up the pairs across words and within speakers. S's voiced durations and gestures are somewhat discrepant, but in fact his utterance of this particular stimulus was acceptable .

Figure 8a would appear quite similar if, instead of tongue height, it were a display of T's and S's pitch traces in the Mandarin MP "āi-ǎi." This tone 1-tone 3 contrast would, when superimposed, show comparable difference patterns. This point is made to emphasize the generality of the "sliding-MATCH" function illustrated here. Since baselines are never plotted, differing ranges of fundamental frequency cannot affect tone MP's; different ranges of algorithm output caused by different vocal tracts are similarly unobservable.

4.3.7.2 Inter-speaker MATCH mode. When phrase or sentence materials are used, the overall shape of the speech gesture as analyzed by the system becomes important, not merely the contrasts between its parts. If the analysis trace is continuous (as would be the pitch trace for the unbroken voicing in "How are you?"), a sliding, intra-speaker MATCH is impossible.

A

B

Fig. 8. CRT Display during the MATCH Function for Two Types of Stimuli. (See text for explanation.)

Figure 8b illustrates the superimposed pitch traces of S and T in the Mandarin example. Note that the two utterances' onsets have been superimposed, and that no further temporal manipulation has been done on the balance of S's trace. The smooth transition from Fig. 7b to Fig. 8b and back again takes an amount of time similar to that used in the sliding MATCH. In matched position, S's (unplotted) pitch baseline is superimposed on T's; the T trace is unmoved; and the S trace is also adjusted horizontally, if needed.

If S and T have differing fundamentals, superimposed pitch plots for similar utterances would be parallel, given the logarithmic plotting of pitch. If tongue-position plots are vertically matched—which, though unlikely, remains a possibility for some types of vowel training—then the same parallelism would obtain for similar vowel pairs spoken by dissimilar vocal tracts. In both of the above cases, S would have to be instructed to disregard any constant offsets between his and T's traces. Over a period of trials, S will get a feeling of what his usual offset is, and it will adapt out perceptually; then, departures from this level will be as clear as discrepancies in the sliding MATCH described above.

Legal operations following MATCH are: HELP, NEXT, STORE, MATCH, PLAY, and FREE.

The MATCH function is the last operation available to S that is relevant to the speech display. In summary: Ss pass to the following stimulus via NEXT: at that point, they must record an unbiased first version of the stimulus via STORE; then, things become less constrained. S may go back to NEXT; he may re-enter a new attempt on the tape loop via STORE; he may MATCH, REPLAY, or enter FREE mode. After RECALL or MATCH, all functions are legal; after RETURN from FREE, all save REPLAY.

Note that nowhere in the present system does the software provide any quantified evaluation of the adequacy of S's imitation of T's trace. Such a capability is beyond the present state of the art. Despite this lack, the richness and interactivity of the audiovisual comparisons that S may make are such as to increase the probability that S's pronunciation will move in the direction of the model. We do not discount the importance of good instructions, but neither do we depreciate the integrative powers of the student.

4.3.8 HELP Button

If at any time S has a problem with the equipment or has a question he cannot answer on his own, he may press the HELP button. This removes whatever is on the screen and replaces it with the legend "TALK." His microphone output is then connected to the outside speaker and to the filter bank. The equipment monitor can then listen to S's question. When he wishes to respond, he need only address his answer to the external microphone, which is also connected to the filter bank. The filter bank is connected alternately to the two microphones, and is switched at the rate of 50 times per second. When the loudness function computed from the monitor mike exceeds some threshold, that signal is passed into the student room, and the display is changed to read "LISTEN." Appropriate time constants are used to insure reasonable voice control by S and the monitor. They can hold a conversation quite easily in this way, without either of them leaving his station, or otherwise interrupting the session. When S is satisfied, he exits via the RETURN button and proceeds at the state just before the interruption.

If the monitor notices, during the course of any session, that S is becoming confused or is in need of some guidance, then he may break into the proceedings himself through striking a key of the

PDP-8E teletype. The program will interpret this input in the same way as S's press of the HELP button. The monitor can interact with S as above and then relinquish control of the system to S.

4.3.9 Privileged Functions

Figures 3a, b, c and d, presented above in Section 4.2, are "frames" from what is terned the "spectral movie." While never shown to S due to their complexity, they are retained in the present software because of their general usefulness in system development and equipment checkout. When the general display program is started in privileged mode, the five central buttons in the button box become functional.

4.3.9.1 ENTER/START. When activated (only legal when the (+) sign is displayed), the standard picture is removed and the first of the 200 frames is displayed. Frames are displayed for 50 msec. each, producing a 1/5th real-time rate. A small cursor, visible below the abscissae in Fig. 3, moves one notch to the right for each new frame. If ENTER/START is depressed during the running of the movie, the frame is reset to #1 and the movie proceeds.

4.3.9.2 HOLD. Halts the movie when running. The number of the current frame is displayed for reference.

4.3.9.3 HOLD/BACK. If running, the movie is halted and the cursor moved one frame back from its present location; if already held, only the latter.

4.3.9.4 HOLD/FORWARD. Opposite logic from above.

4.3.9.5 CONTINUE. When movie is held, allows it to proceed from the present cursor location. Frame number is absent when the movie is running.

4.4 CONCLUSIONS

The design philosophy of the instructional display described above is to present to the student simple visual feedback on a relevant dimension of his pronunciation as a function of time, and to do this in a context of interactive, student-directed audio-visual inspection of the comparison with the speech of a teacher. On the basis of previous research, we have chosen to make the time parameter an explicit one. Single speech functions, selected and tuned to the particular accent problem being emphasized, are plotted in real time as they are extracted from student and teacher speech.

If a teacher were present, beside the student, and if both could interact with the API and with each other, the generality and flexibility of the entire system would of course be greatly increased. A trained teacher could evaluate student speech, both by ear and with reference to the display; could change his speech and/or modify the display in response to particular problems of the student; and, by encouragement and discussion, could mold the articulation of the student. The presence of a trained teacher would also permit the use of more complex displays, with the teacher taking on the burden of pointing out the subtleties that are obvious to the trained eye but obscure to simplistic machine recognition algorithms, or to naive students.

While this might be an ideal arrangement, it would also be an expensive one. Further, it runs counter to the goal of this project, which is to develop an automated system for reducing accent in second languages. The student is therefore alone before the machine, having only his instructions, the display, the buttons, and the monitor to guide him in his work. This constraint has produced a system with robust, simple visual displays and recorded teacher models. The capabilities of the present-day language laboratory have therefore been preserved within the context of a larger, more sophisticated system that can provide visual speech feedback in addition to the standard auditory feedback. We look forward with anticipation to the coming field tests.

5. REFERENCES

Defense Language Institute: Chinese-Mandarin basic course. Defense Language Institute, May, 1964.

Hockett, Charles F. Progressive exercises in Chinese pronunciation. Institute of Far Eastern Languages, Yale University, 1951.

Lado, Robert. Test of aural perception in English for Latin-American students. English Language Institute, University of Michigan, 1957.