Report No. 2185
Job Nos. 11266
11545

INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE

FINAL REPORT, SECTION 1

Task 1: SECOND LANGUAGE LEARNING

30 June 1971

ARPA ORDER NO. 890, Amendment No. 5

Sponsored by the Advanced Research Projects Agency, Department of Defense, under Air Force Office of Scientific Research Contract F44620-67-C-0033

Prepared for:

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

57

DISCLAIMER NOTICE

THIS DOCUMENT IS THE BEST QUALITY AVAILABLE.

COPY FURNISHED CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

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 Final

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

Daniel N. Kalikow

6 REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
30 June 1971	55	4

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
F44620-67-C-0033	
b. PROJECT NO. 890	
c. 61101D	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d. 681313	AFOSR - TR - 71 - 2789

10 DISTRIBUTION STATEMENT

Approved for public release;
distribution unlimited.

11 SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
TECH, OTHER	AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (NL) 1400 WILSON BLVD ARLINGTON, VIRGINIA 22209

13 ABSTRACT

Two experiments testing the effectiveness of the Automated Pronunciation Instructor (API) were carried out with Spanish-speaking students attempting to learn English. Experiment I tested the computer-generated display of the student's tongue position during stressed vowels. The experimental design described in earlier reports was followed, save that additional pronunciation displays - for reduced vowels and aspiration of initial consonants - were not included. New experimental procedures for evaluating the extent of pronunciation changes were developed and applied. Experiment II tested the effectiveness of the additional two displays. Highly significant training effects were obtained for all three displays, despite the presence of variability in the data. This report concludes with a general discussion of proposed changes to future versions of the API system.

DD FORM 1 NOV 68 1473

UNCLASSIFIED

Security Classification

Report No. 2185 Bolt Beranek and Newman Inc.

INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE

FINAL REPORT, SECTION 1

Task 1: SECOND LANGUAGE LEARNING

30 June 1971

by

Daniel N. Kalikow

ARPA Order No. 890, Amendment No. 5
Sponsored by the Advanced Research Projects Agency,
Department of Defense, under Air Force Office of
Scientific Research Contract F44620-67-C-0033

Prepared for

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

TABLE OF CONTENTS

FINAL TECHNICAL REPORT

ARPA Order No. 890, Amendment No. 5
Program Code No. 9D20
Contractor: Bolt Beranek and Newman Inc.
Effective Date of Contract: 1 November 1966
Contract Expiration Date: 30 June 1971
Amount of Contract: $804,896.00
Contract No. F44620-67-C-0033
Principal Investigators: John A. Swets
Mario C. Grignetti
Wallace Feurzeig
M. Ross Quillian
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 develop a computer-based system for automated instruction in the acquisition of the new speech sounds of second languages, and to ascertain the efficacy of the approach through experimental tests.

2. General Methodology

Laboratory experiments.

3. Technical Results

Two experiments testing the effectiveness of the Automated Pronunciation Instructor (API) were carried out with Spanish-speaking students attempting to learn English. Experiment I tested the computer-generated display of the student's tongue position during stressed vowels. The experimental design described in earlier reports was followed, save that additional pronunciation displays - for reduced vowels and aspiration of initial consonants - were not included. New experimental procedures for evaluating the extent of pronunciation changes were developed and applied. Experiment II tested the effectiveness of the additional two displays. Highly significant training effects were obtained for all three displays, despite the presence of variability in the data. This report concludes with a general discussion of proposed changes to future versions of the API system.

4. Department of Defense Implications

Language schools for 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.

1. PREFACE

At its inception in 1966, this contract was devoted solely to the one area of second-language learning. Later amendments have added three more tasks: Models of Man-Computer Interaction; Programming Languages as a Tool for Cognitive Research; and Studies of Human Memory and Language Processing. The present contract was scheduled for termination on 31 December 1970, but the final reporting date was changed to 30 June 1971, to allow completion of data analysis in the various tasks.

Due to the amount of information to be presented in the Final Report, we have bound it in four Sections, one for each task. In addition to a copy of this page, each Section contains an appropriate subset of the documentation data required for the report: a contract-information page, a summary sheet for the particular task at hand, and a DD form 1473 for document control.

2. INTRODUCTION

We describe here the conclusion of four years of research on, and development of, a computer-based Automated Pronunciation Instructor (API) system for aiding students in learning second languages.[1] Previous technical reports have described the phonological research on the specific problems encountered in the Spanish-English language pair, a teaching environment that had been picked as the paradigm for the work. The process of hardware and software evolution through two successive realizations of the system has been similarly described. The process of integration of these two channels of activity had proceeded so well by June 1970 that the first formal evaluation of the effectiveness of the prototype API was appropriate. An evaluation experiment was designed and begun, and data were collected through the final months of the present contract. BBN has since been awarded a new contract for continued system development and evaluation within the context of the Defense Language Institute schools.

The overall design of the first of the evaluation experiments was presented in Semiannual Technical Report No. 7. There was a truncation of the planned scale of this first test; only the previously-described vowel tongue-position display was evaluated. In that work, a new and sensitive means of accent analysis was devised and used. The second experiment dealt with two newer displays: reduced-vowel tongue position, and aspiration-voice onset.

---

[1]The author gratefully acknowledges the contributions of the other members of the project team: Dennis H. Klatt, Kenneth N. Stevens, John A. Swets, and Douglas W. Dodds. He thanks Barbara A. Noel for monitoring the experiments, and Karl S. Pearsons and Sanford A. Fidell for aiding the data analysis.

The following section of the report summarizes these experiments, including mention of the most important aspects of display generation, analysis technique, and outcome. This summary concludes with some remarks on what the evaluations have taught us, and on the implications of this work for the new versions of the API now under construction. Should the reader wish additional detail on some facets of the work, the summary will direct his attention to the appropriate section of the appendix.

3. OVERVIEW[1]

3.1 EXPERIMENT I

3.1.1 Method

Seventeen women, born in Latin America and speaking Spanish as their native tongue, were interviewed. Tape recordings were made of their utterances of English words in the standardized format of the testing procedure described in section 2.6 of Semiannual Technical Report No. 7. In brief, this involved their reading and speaking a set of English words, one at a time, after hearing a recorded version of each word as spoken by a native English talker. The model pronunciation was used to alleviate orthographic contamination of the potential student's speech. Mimicry, the other side of the coin, was minimized by the institution of a forced 4-second time delay between model playback and recorded utterance.

Each woman's speech was rated, with higher ratings going to those having more accent in the production of vowels. The highest ten persons were selected as subjects (Ss), and they were assigned to either the Experimental or Control treatment groups according to a matching process (see Appendix 4.1, Experiment I Subject Selection Procedure, for further detail).

The training procedures previously outlined were followed. Twice a week, each S worked for 45 minutes with one of four sets of 12 monosyllabic English words, each containing one of the 12

---

[1]This section is not to be construed as a quick guide to the appendices. It is the sole location within the report where an integrated picture is presented. Where backup detail is deemed necessary, the reader is referred to the appropriate section of the appendix. As far as possible, each appendix is self-contained.

vowels for which the system provided tongue-position feedback (for the Experimental Ss). Control Ss spent the same amount of time with the API, but the system provided no tongue-position feedback via the CRT display.

Eight of the original Ss completed 16 training sessions. One control S was lost when family illness forced her to return to South America; another completed only 13 training sessions before unexpected travel plans forced us to terminate training and post-test her, rather than lose her data. The remaining Ss were post-tested later, following the identical test-day format outlined earlier. Retention-testing was carried out, for all 9 Ss, a minimum of four weeks after post-testing. There was no intervening contact with the Ss.

3.1.2 Results

The data from the experiment were of three types: (1) audio tape recordings of Ss' speech during the three testing days of the experiment; (2) audio tape recordings of Ss' speech made automatically during selected normal training sessions; and (3) punched paper tapes produced after each training session, quantifying the manner in which the Ss distributed their efforts among the 12 training words.

The paper-tape data were inspected to determine whether the amount of activity across the 12 trained vowels different significantly between the two treatment groups. Several two-factor mixed analyses of variance (Lindquist, 1953, pp. 266-273; Winer, 1962, pp 302-312) were performed on summarized data derived from all training sessions from each S, a different analysis for each type of activity (pressing STORE button, RECALL button, etc.). A pattern

of effort was demonstrated; i.e., activity was not homogeneous across vowels. However, no differential effect of treatment on the pattern was found.

Audio data in (2) above, from the normal training sessions, were used in the investigation of a warm-up and reminiscence effect that had been noted during informal observation of the training-session utterances of the Ss. This effect is secondary to the main point of the experiment, and will be mentioned further only in the detailed coverage provided in the appendix (see Appendix 4.2: Experiment I Data Analysis).

Audio data in (1) above, arising from post- and retention-testing sessions, were edited and rearranged into a standard order. Since only the vowel display had been used, only the 24 vowel words were extracted from the recordings made on the three testing days. Each S's course of training was thus succinctly represented by three segments of magnetic tape containing the same set of 24 monosyllabic vowel words, recorded at three points in time. These were the primary data from the experiment.

An elaborate pair-comparison process was used to inspect the Ss' utterances for changes in adequacy occurring in the course of the training. A subjective rating procedure implemented by a computer-controlled system for tape transport and response recording system allowed a large number of judges (Js) to indicate their preference between many randomly ordered pairs of utterances. Each utterance pair was preceded by the playback of a proper pronunciation of the word being attempted by the S. The pairs themselves were drawn from two of a given S's three recordings of a given word on the three testing days. The Js were of course uninformed of the actual time relationship between the word pairs, and their instructions were to indicate "which of the two student

pronunciations contain(ed) the vowel that sounded more like the standard pronunciation." The measure of success was the number of times the Js' preferred utterance was actually obtained later in training. Thirty-two Js responded to a total of 120 trials from each of 9 Ss' recordings. Each run contained catch trials to measure reliability, and the three utterances of each of the 24 words were intercompared exhaustively, producing validity and transitivity information. A total of 34,560 responses was collected and processed.

Conventional two-way mixed analyses of variance did not reveal any consistent advantage in amount of improvement shown by the experimental Ss. This was caused, in some instances, by ceiling-effect perturbations from a strong overall training effect, and in other instances, by excessively strong response variability. The analysis of variance approach was therefore abandoned in favor of more global summary statistics, designed to consolidate the data to minimize variability and enhance sensitivity to specific effects.

This process allowed the observation of highly significant training effects in both experimental and control Ss. A representative summary statistic demonstrating this effect is the group pre-post pair preference percentage, giving the percentage of the time that judges chose a post-training utterance as preferable to an utterance of the same word recorded at pretesting, for all words spoken by all Ss within a given treatment group. This statistic takes on the values 62.1% for experimental Ss, and 61.0% for the control Ss. Considered against the null hypothesis of no training effect and evluated with the normal approximation to the binomial, these values convert to standard scores of 14.9 and 12.2, respectively, yielding a differential treatment effect of 2.7

standard score units in favor of the experimental group. Pre-retention preference percentages were on the same order of magnitude, and the differential advantage of the experimental Ss increased to 3.0 standard score units. The third side of the judgmental triangle, post-retention, complemented the first two statistics and elucidated changes occurring within the no-treatment interval. Overall, these group-level comparisons across words and within a given pair of testing days showed a consistent advantage for the experimental treatment, over a strong traning baseline of the control treatment. Though both experimental and control Ss' retention-day performance was significantly better than their pretest performance, it is worthy of note that the experimental Ss continued to improve during the retention interval, while the control Ss worsened slightly. This consolidation over the one-month no-treatment interval may possibly indicate a particular strength of the API.

An even more powerful way of inspecting the data follows from the notion of response transitivity. Each judge responded to three pairs of each word and subject. There are 8 possible judgment triads that can arise by chance. Two of these are reflections of intransitive stimulus orderings, and the other 6 order the pre, post, and retention utterances in their possible permutations. Of these, two place the pretest utterance at the bottom of the preference continuum. Therefore, the expected rate of occurrence of judgment triads meeting this latter criterion is 2/8, or 25% by chance.

Occurrences of such stimulus orderings were tallied for all words spoken by Ss in a treatment group, across all Js. For the experimental Ss, 42.5% of the triads met the criterion; for the controls, 41%. Translating this into standard scores measuring deviations from the chance expectation of 25%, we find the controls

manifesting a strong training effect: 20.4 standard deviations from expectation. However, the experimental Ss produced data 4.7 units more removed from expectation, giving evidence in favor of an advantage attributable to the full feedback capabilities of the API system.

3.1.3 Discussion

Two major observations are appropriate at this time. First: the present control treatment is too similar in rigor to the experimental treatment, making it difficult to extract effects attributable solely to the visual feedback. The API system should be evauated by comparison with state-of-the-art procedures for pronunciation instruction, and not by comparison with a stripped-down version of itself. Second: future evaluation efforts should include measurement of generalization of trained speech sounds into new words not specifically trained, and into connected speech, despite the latter's multidimensional nature.

Despite the above reservations, this experiment demonstrates that the vowel pronunciation of Spanish-speaking Ss is significantly improved through training with the API system. This improvement with training was more marked in Ss exposed to the computer-generated acoustic analysis and visual feedback. Byproducts of the experiment were the development and implementation of a powerful evaluation procedure of general usefulness in scaling training differences between similar utterances, and general experience to be applied toward the design and execution of future display-evaluation experiments.

3.2 EXPERIMENT II

Since the first experiment had dealt only with the vowel tongue-position display, a second experiment was executed to evaluate the additional displays that had been produced.

3.2.1 Reduced-Vowel Tongue-Position Display (RVTPD)

In English words, vowels that appear in unstressed syllables must be "reduced." That is, they must be short in duration and they must take on the spectral quality of the schwa vowel. Since this sound is not present in Spanish, our students tend to render the word "dif-fuh-cult" as "dif-fee-cult."

The first step in the display is the isolation of the relevant syllable from the multi-syllabic training word. This is accomplished by an algorithm that searches for time maxima and minima through a function produced by summing the outputs of filters 2 and 3 for each time sample of digitized speech (see Appendix 4.3 on the RVTPD for specifics on the algorithm). These filters were chosen because they indicate the low-frequency energy characteristic of voicing in vowels. Once the samples produced by the relevant syllable have been isolated, the analysis of the vowel nucleus proceeds in a way that is virtually identical to that used above for stressed vowels in monosyllabic words. The same view of the mouth is presented, with a target of appropriate size and location for the schwa vowel.

3.2.2 Aspiration-Voice Onset Display (AVOD)

In the initial aspirated stops /p, t, k/, Spanish speakers must learn to delay voicing onset with respect to stop release, and they must learn the glottal gesture required to produce aspiration during this interval. In other words, in pronouncing a word

like "team," a common error is to begin the vibration of the vocal cords too early after the start of the utterance, and therefore to minimize the required 'puff of air' between the t and the e. For an English speaker, the result is difficult to discriminate from "deem." There are three parts to the display for this problem.

1. Stop release time is determined by the discovery of a time sample containing a filter in the range from filter 4 through 19 whose contents have increased more than a threshold amount with respect to the immediately preceding sample from that filter. This is therefore a procedure sensitive to sudden amplitude changes such as occur at stop release.

2. The presence of intense low-frequency energy indicates the onset of voicing. The earliest time in the utterance is found such that the output of filter 2 exceeds a threshold. This threshold is normalized to the maximum value of that filter's output during the entire word, to compensate for the recording level of the utterance. If this algorithm is satisfied at an earlier time than (1) above (as in the error "deem"), stop release is made equal to voice-onset time.

3. Aspiration intensity is computed for the samples lying between the two above points in time. This is given simply by the summed activity in filters 14 through 16, since this frequency region is activated by aspirated sounds. The aspiration function is normalized before display to compensate for differences in recording level.

For the various training words used, nominal values for voice-onset time and aspiration intensity were obtained from English speakers, to be used as target values for the students. Figure 1 shows CRT output following a successful attempt at the word "team."

NOT REPRODUCIBLE

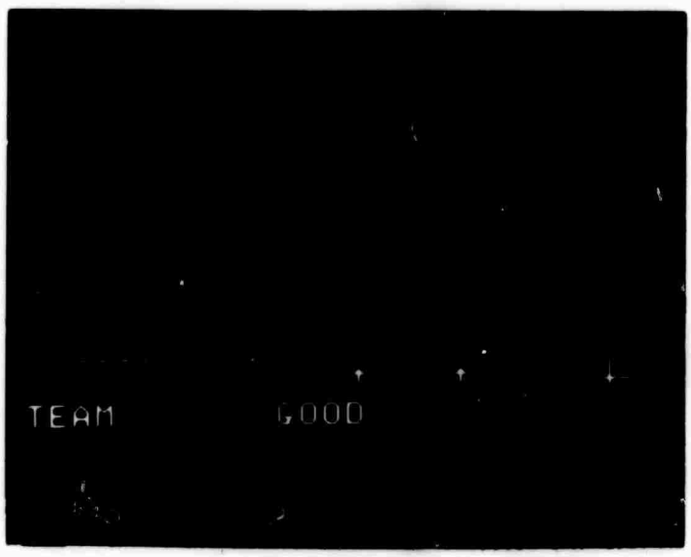

Figure 1 Aspiration-Voice Onset Display

This is a plot of two variables as a function of time, which moves to the right. The vertical line is placed at the point of stop release. The widely spaced dots indicate aspiration intensity at successive 10-msec intervals. The horizontal row of more closely spaced dots indicates nominal aspiration intensity. At least one time sample of aspiration must meet or exceed this line in a correct utterance. The abscissa line is interrupted at the point of stop release, and its plotting is resumed at voice-onset time. The region of acceptable voice-onset times is delimited by the arrows.

If one or the other of the display criteria is not met, the appropriate indicating section of the display (dot row or arrows) is made to blink intermittently, calling the student's attention

to the type of error made. If both criteria are met, the display is steady and the word "GOOD" appears, further reinforcing the student.

3.2.3 Method

In addition to the new software, a high-fidelity head-mounted microphone was used in this experiment. This standardized the mouth-to-microphone distance, a necessary feature for proper action of the aspiration display.

Several potential Ss were pretested with procedures essentially similar to those used in Experiment I. Thirty-six words were recorded, half for each of the two displays. Twelve of the words were the training stimuli for each display, and the remaining six were uttered only on the testing days. Accent-rating procedures were used to determine whether a potential S demonstrated the pronunciation problem relevant to the display. Four acceptable Ss were selected, three of whom had served in the preceding experiment. No control treatment was employed in this experiment, for the reason stated above. The specification of a proper control treatment was left as a problem for the future; in the meantime, quantification of the API's full capabilities was the primary concern.

Six training sessions, two per week, were administered. Each session included 25 minutes exposure to each display, with a 5-minute break between sections. At the time when the seventh training session would have been executed, all Ss were post-tested in the standard way. The basic data for Experiment II were prepared in the form of two segments of tape for each of the 4 Ss. Each segment contained the standardized utterances of the 36 test words, spoken before and after training.

3.2.4 Results and Discussion

Ten judges indicated their preferences for either the pre- or post-test version of each of the 36 words spoken by the 4 Ss The Js had varying degrees of knowledge of the intent of the experiment, but all had been instructed to prefer those utterances having the better pronunciation of the feature of interest, and to attempt, as far as possible, to discount others.

The definition of a training effect is identical with the previous one: S is found to have improved her accent on a given word if her POST utterance is judged preferable to her PRE utterance at a level reliably above chance. A training effect is a reflection of the fact that Js can reliably order utterances in a way that is congruent with the application of training.

What follows are some summary statistics derived from the entire group of 4 Ss (for more detail on the analysis, see Appendix 4.4). For the reduced-vowel display, on the trials in which Js expressed their preferences between words actually trained, 64.4% of their judgments were in favor of the post-test version of the words. For words not specifically trained, the preference rate dropped to 55.8% in favor of words produced after training. The response level across all words, 61.5%, produced a standard score of 6.2 as evaluated with the normal approximation to the binomial, and it was therefore highly significant.

For the aspiration display, the results were more striking. The preference rate for post-test versions of trained words was 85.6%; for generalization words, 86.6%; and overall, it was 85.9%, for a standard score of 19.3.

The RVTPD thus demonstrated approximately the same power as the VTPD examined in Experiment I, as evidenced by the appropriate comparisons of pre-post pairs of trained words (62% for VTPD, 64.4% for RVTPD); and the AVOD display, with its totally different display structure and target articulation, was substantially more effective.

3.3 GENERAL DISCUSSION

3.3.1 Experimental Design

The basic reason motivating the choice of the Spanish-English language pair as the medium for the first system evaluation was our familiarity with English phonology and the relatively simple pronunciation problems involved in Spanish speakers learning English. It continues to be a useful language pair. However, it did force the use of a subject pool having considerable experience with English, with consequent overlearning of incorrect pronunciations. Ideally, the API should be used when the student begins contact with the second language. In that context, it should aid in acquiring new sounds faster and correctly. This problem will be somewhat minimized in the coming field evaluation with Spanish-speaking DLI students, since their previous contacts with English will be more clearly specified. The problem will be eliminated for the English-Mandarin Chinese language pair.

Despite the above difficulties, there is no doubt that the API has produced significant improvements in the rendition of certain English phonemes by Spanish-speaking Ss. However, future evaluation experiments should be less concerned with fine-grained measurements and more concerned with the demonstration of effects in connected speech. This is not a simple task, since the basic mode of operation of the API is to work with limited speech samples,

and since the problems of evaluation are manifold. Some suggested additions to the evaluation program are:

A. Recording samples of student speech at various times during experiments, and their rating by an experienced panel. This might easily be done within the standard context of Defense Language Institute student-evaluation procedures.

B. On the assumption that training with these displays improves the ability of the Ss to make auditory discriminations between correct and accented utte ances of the various phonemes for which displays have been developed, in order to produce improved versions of those phonemes, various perceptual tasks might be attempted. After procedures presented by Lado (1961), auditory discrimination tasks could be administered before and after training, to determine whether the API aids that faculty while aiding the production capabilities of the Ss.

3.3.2 Hardware Changes

Within an acceptably short time, all Ss learned to make satisfactory use of the API, with the help of the instructions, monitor, and display software itself. Still, there were residual difficulties that placed unnecessary cognitive loads on the Ss, diverting their attention from the task at hand, as follows:

A. Language Master unit. The stack of tape-recording cards, containing teacher versions of the training words to be entered by S onto the tape loop, had to be kept in order and entered correctly if the software was to keep up with the word being studied. This error source will be replaced with a completely automatic (and higher fidelity) means of teacher recording storage.

B. Button Box. Each S was provided with a written guide to the functions performed by each of the 12 buttons before her, but the buttons could not be labelled individually. The new system will have a custom-designed button array built into a desktop, containing color-coded buttons, each of which can be internally lighted by the software. Thus, S will be visually informed of the options at each choice point. Sufficient space around the buttons will allow a key-word description of each button's function.

C. Error Diagnostics. The improved central processing unit of the Mark II API (a PDP8/E instead of the PDP8/L used in the Mark I) will make it possible for the system to be more interactive and informative when S makes errors. The Mark I was limited in core storage to 4K, barely sufficient for minimal diagnostic action in the VTPD and RVTPD software. We plan to employ 8K of faster memory in the coming system, along with hardware multiply/divide and improved analog-to-digital capabilities. The Mark II will thus be able to inform S more fully about needed articulatory corrections, and will pinpoint more specifically any missteps in system use.

3.3.3 Changes in Teaching Strategy

The time-plotting nature of the aspiration display may well be a strong factor in its strength relative to the tongue-position displays. The latter displays show time as the parameter connecting successive position points on the CRT "map" of the mouth; the aspiration-voice onset display shows two variables plotted explicitly as functions of a common time line. Work is now underway on the production of time-plotting tongue position displays, in an attempt to combine the best features of our previous work into a new display type.

We are also in the process of adding a totally new teaching procedure to the API system. The use of minimal-pair utterances was suggested by interested observers from DLI, and the potential utility of the approach is confirmed by an inspection of the current literature on pronunciation-teaching procedures. A minimal pair is two words in the target language, with all phonemes identical save one. Their use in pronuciation teaching is simple: S produces his version of the pair, and some evaluator (teacher or API) produces feedback on the adequacy of the distinction between the contrasted phonemes. The new system will incorporate this more extensive utterance mode. Its larger memory capacity and its capability for automated storage and retrieval of recordings of teacher versions of minimal pairs will make possible entirely new modes of teaching.

The final area of pedagogical improvement is related to all of the preceding proposed improvements, and is in a sense the most basic. Through all our work to date, we have provided Ss with various CRT targets. When the feedback met these criteria, S was told (explicitly or implicitly) that his pronunciation was correct. However, the type and range of stimulus words for which feedback could be generated has always been limited by the necessity of "tuning" the targets for each word. This was a process of empirical extraction of the invariant properties of correct utterances, and their automated specification by the software. If this bottleneck can be eliminated, the way will be opened for large stimulus sets and generality of training stimuli. The prime unused resource of the Mark I API is the world's greatest pattern-recognition machine: the mind of the student. It is capable, given the proper visual input, of abstracting many more aspects of correct utterances than can ever be automatically specified, given large numbers of instances from which to generalize. We plan to use this capability

in new display modes within the Mark II, in which the system will simulate what a language teacher does in pronunciation instruction. He does not simply repeat the training material ad nauseam, and he does not mimic the student's mispronunciation. Rather, he gives verbal descriptions of how he positions his own articulators in addition to telling the student in what way he should reposition his. The automated source of teacher recordings introduces the capability for software analysis of teacher utterances, in the same terms as the student's utterances are analyzed. The student will thus be able to use as a template an actual teacher's utterance, whose salient characteristics are emphasized by the software in a manner identical to the analysis of his own voice. This enhances the simulation of the presence of the teacher. Since little "tuning" is needed because of the concurrent analysis, many different teacher utterances relevant to a given phonological problem may be presented in a single session. Current plans are to implement teacher analysis within the framework of minimal-pair training material, and using the time-plotting approach mentioned above. The coming version of the API will receive its most stringent test in this type of display, since it will incorporate our most advanced thinking.

In conclusion: our past efforts on this task have been rewarded by the successful production and testing of a prototype model of an Automated Pronunciation Instructor. The development and evaluation of the system has not been without some problems in concept and execution, but the basic value of the approach has been demonstrated. The course of our work has generated new information and ideas to such an extent that we are confident that the next iteration of the system will be considerably closer to the breakthrough in pronunciation instruction so sorely needed. We look forward with anticipation to the transfer of this new technology to the field.

APPENDIX 4.1: EXPERIMENT I SUBJECT SELECTION PROCEDURE

As indicated in Section 2.6 of **Semiannual Technical Report No. 7**, the only valid method for S selection was the full pretesting procedure outlined therein. A total of 17 Spanish-speaking women were tested, and the tape recordings of their responses in this standardized milieu were edited and rearranged.

In each S's final pretest tape, there was but one version of each of the critical words: that version spoken by S after having read the Language Master card into the system, after having heard the teacher's voice speaking the word, and after waiting through the "countdown" displayed on the CRT. Occasionally, more than one utterance of a word occurred before the final version was accepted by S, L, and the machine. The incidence of multiple repeats declined through each potential S's pretesting session. The edited tape recording contained the critical words in a standard order, to facilitate subsequent rating procedures.

Rating procedures were employed to effect selection of Ss and to assign them to treatment groups. Each of the 17 interview tapes was played for a panel of experienced judges, who assigned to each of the utterances of the Ss a number ranging from 1 to 4, the latter indicating extreme accent. Overall accent scores for each S were computed according to a weighted averaging procedure, with the most weight going to the pronunciation of the 24 critical vowel words.

Potential Ss were then ranked in order of decreasing weighted accentedness score, and the highest ten were selected as actual Ss. They were assigned to either the experimental or control groups according to two criteria: (a) one member of <u>each successive pair</u> of Ss in the accentedness list must go into the experimental group; and (b) the results of this pairwise assignment should produce two samples of speakers with roughly the same histories of exposure to English. This procedure resulted in an experimental group with average values of 6.0 and 5.7 years for the study of English and residence in the United States, and in a control group which averaged 4.6 and 7.3 years, respectively. There are no <u>a priori</u> data on which of these poorly-specified factors has more influence on the acquisition of accent-free speech, and so assignment to the two groups was adjusted to achieve a fair balance between the factors, keeping overall accentedness matched.

The main strength of this selection procedure was that it was done in terms of the behavior to be tested. A group of speakers was found whose pretest-day utterances were evaluated and found to be sufficiently accented. A known procedure was then used to form the two treatment groups. The data upon which selection as S was based were the same data that were to serve as a baseline for the evaluation of treatment effects: i.e., the edited tape of pretest utterances. Since the system was designed to aid the production of the same type of utterances, the S-selection procedure was closely fitted to the capabilities of the machine they were to use.

APPENDIX 4.2: EXPERIMENT I DATA ANALYSIS

Theory. We chose to investigate the "warm-up effect" mentioned above by editing out, from the training session history tapes, the last successful utterance of each of the critical words made during the course of training. Since the 24 vowel words were spread across the four training lists, the construction of the end-of-session (EOS) tape for each S involved the selection of the words from the history tapes of four different training sessions, as far advanced in training as possible. The EOS tape thus formed the fourth segment of audio information from each S available for later analysis.

The central data of the experiment were contained within the set of 36 sections of tape, four sections from each S. Each section contained the 24 critical words as uttered by that S in a standardized recording format: the PREtest, POSTtest, and RETention test tapes in the critical day milieu, and the EOS tape as gathered from representative normal training sessions for that S. We looked to these 24x9x4, or 864, words, to determine improvement with training and treatment.

The primary variable investigated in the experiment was training, whose representation in the data is the time separating the recordings of the 24 words. Therefore, pairs of words to be compared should all involve the same word spoken by the same S at different points in time. From the four versions available for each of the 24 words and nine Ss, six different pairs could be constructed, with order not considered. Four of those six were considered essential:

PRE-POST (mnemonic PEPO): compare a word as spoken before and immediately after training;

PRE-RETENTION (PERE); compare baseline utterance with same word spoken after a no-treatment interval, to test whether it has undergone any long-term changes;

POST-RETENTION (PORE): compare utterances spoken immedi tely before and after a retention interval.

Consider a given word spoken by a single S at the three points in time. If a judge is asked to state his preference for one member of each of the three above pairs, his responses should be mutually consistent. Taken together, these three pairs form a judgmental loop, about which more will be said below. The final pair judged essential was:

PRE-EOS (PEEO): compare baseline utterance with end-of-session utterance.

While it is possible to compare EOS with POST and RENTENTION, the warmup effect could be just as well-quantified by a single comparison, which would place it within the framework erected by the first three pairs.

Since the pairs were to be drawn from one S at a time, that implied the presentation of a minimum of 96 pairs to cover the data of a single S, not counting the administration of some kind of check on the reliability of the responses of the judges. Further, to avoid sequential dependencies, more than one order of stimuli should be administered to the panel of judges.

Apparatus and Procedure

We report here an analysis method for these data which uses the old pair-comparison paradigm in a new setting for accent rating. The system is efficient in terms of speed of data acquisition and mathematically powerful in terms of the number of statistical questions answerable with the data produced. The procedure was developed jointly by the author and the staff of the Psychoacoustics Department at BBN's Los Angeles office. At that facility, there exists a computer system, interfaced with a six-channel tape-cartridge playback machine. Audio material on the cartridges may be presented to groups of judges (Js) sitting in an anechoic chamber. Each of the four Js the chamber can accomodate at any one time has before him a button-box connected to the computer, through which he can record his responses. Figure 2 presents a block diagram of the system.

The four sections of tape from each S were placed in four separate cartridges. Each of the 36 cartridges from all the Ss contained the critical words in the same order. A final cartridge was prepared for the judgment tests. This contained the model pronunciations—recorded from the actual Language Master cards—that the Ss were constantly trying to approximate.

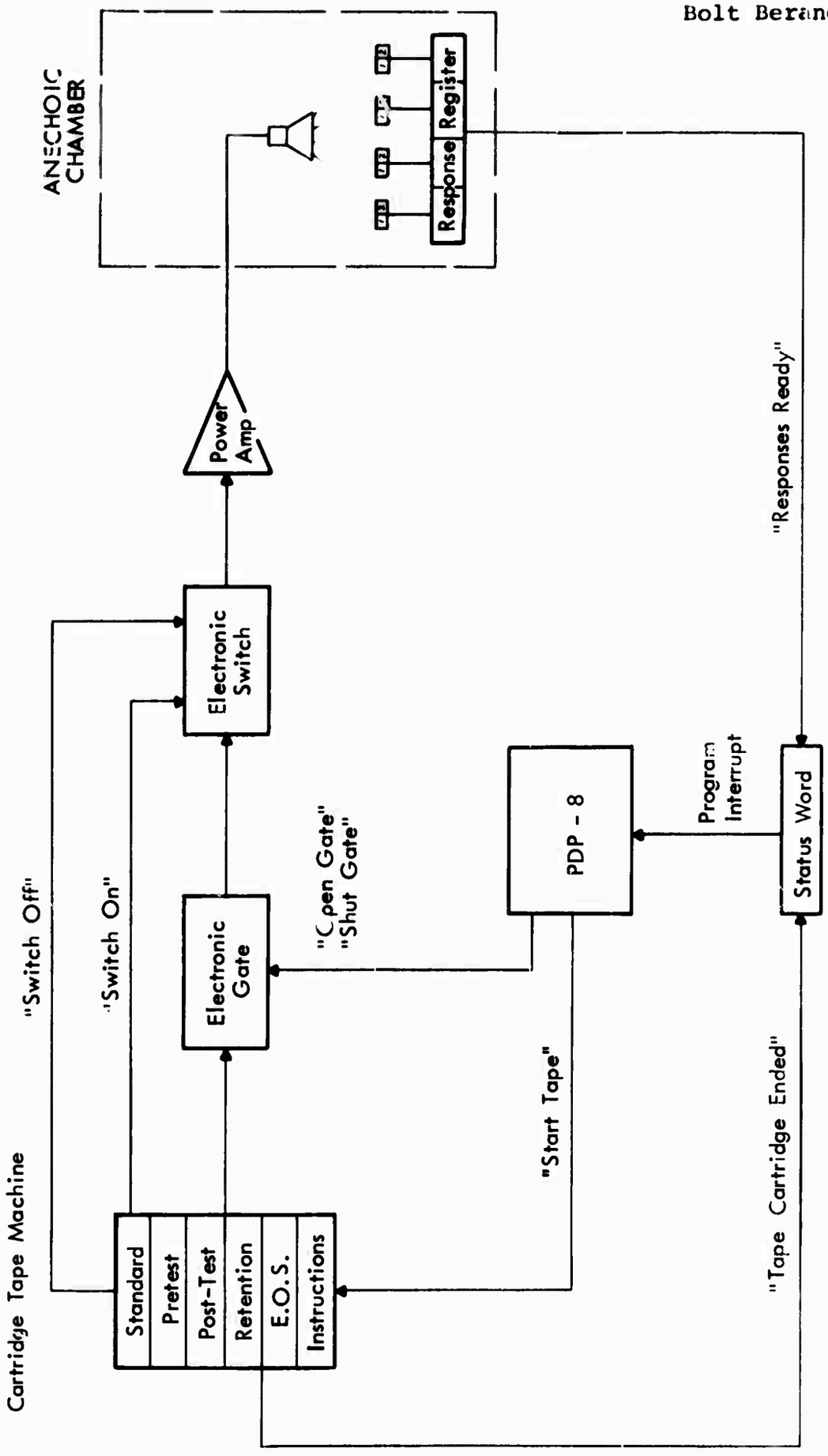

Fig. 2. BBN Experimental Accent Analysis Facility

The analysis was carried out in several sections. Eight different groups of Js participated, replicating it eight times. Js were college students having normal hearing, and were paid $2.50/hr. They sat facing a loudspeaker at a distance of approximately four feet, listening to the speech at comfortable listening levels (i.e., in the range of 70-75 dBA). For each group of Js, the analysis was divided into nine sections, each section being devoted to the utterances of one of the nine Ss. Each section lasted about 15-20 minutes, and was separated from the next by a short break. No more than five sections were performed in a single day by any group of Js. Each section was administered by the computer system in a standard manner. Because of the many degrees of freedom in the algorithm to be described, and because of the automated nature of the procedure, the same stimulus order was never administered twice.

During any one section of the analysis sessions five cartridges were available for playback by the computer system. The cartridge with the teacher's version of the critical words was always present. The four cartridges containing the utterances of the S being judged were inserted for the duration of the section. All were positioned such that the same word was available at each playback head. The starting word was rotated across Ss and groups.

Upon entering the chamber, each group of Js heard a recording of their instructions. These will be presented below, as the best description of the situation facing the Js.

> You are about to evaluate the results of an experiment in which native Spanish speakers were attempting to improve their pronunciation of certain English vowels. Your answers will help us to decide how successful our teaching procedures were. You will be asked to push a button corresponding to the best student pronunciation of an English vowel sound. On each trial you will hear three English words in succession. The first word will be heard in a standard English pronunciation. The next two words will be spoken by one of the students. Your job is to decide which of the two student pronunciations of the vowel sound was more like the standard English pronunciation that you heard first. Listen to the

following sample trial. You will hear first the standard English pronunciation, then two attempts by the student to pronounce the same vowel sound. (Sample heard here.) Here is another sample trial, listen carefully for the three pronunciations. (Another sample is heard here.) Please note that your judgment of which student pronunciation was more like the standard pronunciation should be restricted to the vowel sounds only. During the test you will push one of your two response buttons to tell us which of the two student pronunciations contain the vowel which sounded more like the standard pronunciation. Try to ignore any other extraneous speech sounds.

Of course there are no right or wrong answers. You are the jury, so consider your answers carefully throughout the course of the experiment. Since there are no right or wrong answers, you must make up your own mind which of the two student pronunciations of the vowel sounds was more like the standard pronunciation. Pay no attention to which buttons the other members of the jury happen to be pushing on any given trial.

To help you keep track of the various speech sounds the response buttons in front of you will light up to indicate which of the two student pronunciations is being heard at the moment. As soon as the light goes out in button two, you may press either button 1 or button 2 to tell us your decision. The light in the response button will go on momentarily as you push it. After all the members of the jury have indicated their decisions the button you pushed will light up for about a second before the next trial starts. The light will go out when the next set of sounds is about to be heard.

You are encouraged to make your decisions and push the appropriate button as quickly as you feel you can make a reliable judgment. You will participate in nine sessions, each of which will last about 15 minutes. You should make your decisions as quickly as is convenient for you in order to hasten the completion of each experimental session. The operator in the next room will be listening to you and watching you on the television monitor throughout. Address any questions you may have to the operator only. If you have any questions about the procedure, you may ask the operator now. When you are ready to begin this session, push one of your response buttons.

During the section, 120 triads were played for the Js. The five cartridges were rotated five times during the courses of the section, though each word on each cartridge might not have been played five times in total. The first member of each of the 120 trials was the teacher's version of the word to be spoken by S. It set the standard for the Js' responses to the following pair. The teacher tape always played in the first position, but only two of the remaining four tapes were actually played for the Js for the trial for a single word; the other two were positioned silently in readiness for the next trial. Within a triad, the inter-word interval was one-half second. Upon receipt of all Ss' responses to each triad, a fresh set of three words was heard after a 1.7-second interval.

Cartridges were selected for playing on a given trial by an elaborate randomized design. The four basic comparisons of interest—PEPO, PEEO, PERE, and PORE—were heard at least once for each of the 24 words. The actual order of presentation of the test words within a given pair on a given trial was random; i.e., in the PEPO comparisons, POST precede PRE about half the time. Any successive group of 24 trials was quite heterogeneous in terms of the stimulus pair being administered.

Thus far, the contents of 96 of the 120 trials have been discussed. The additional 24 trials were used to check on the consistency of the Js. Since there are no right or wrong answers, the final arbiter of validity is consistency. Therefore, on 24 of the trials, selected randomly beginning after the first pass through the 24 words, the system marked a particular stimulus pair previously administered to the Js for replication. Half of the replication trials were identical to their predecessors, and half were presented in the opposite order.

The computer stored a representation of the stimulus pairs and the orders in which they were administered, and it also stored the responses of each of the four Js within a given section. During the break between sections, it produced paper tapes containing that information along with identification of the Js and of the S

whose speech had been evaluated. When the data of the analysis experiment were complete, each of the 32 Js had produced 120 responses for each of the nine Ss in the source experiment, and there were thus 9x32, or 288, response matrices on paper tape, each separately identifiable. The total number of responses gathered was 288x120, or 34,560. The tapes were sent to BBN-Cambridge for further treatment on a larger computer.

Data analysis. Virtually all subsequent manipulation and analysis of the data was performed under program control. we will not describe the total analysis performed, nor will we describe the basic processes used to produced the statistics presented. The following is a brief characterization of the principles of the analysis process.

Each of the nine Ss spoke 24 words at four points in time. Restricting our attention to just one of those quadruplets of words, we find that it was presented for judgment in the form of four pairs: PEPO, PEEO, PERE, and PORE. Thirty-two different Js responded to those four pairs, though the specific orders in which they were presented differed only between different groups of four Js. The original data were transformed in such a way that the response of a given J to a PEPO trial was standard, with a "2" indicating preference for POST over PRE, regardless of the order in which the pair had been heard.

Each critical word was also tested with a fifth judgment pair, the replication trial. The response of the Js on that trial was also categorized and standardized. Half of the replication trials were in the same order as had been heard previously (mnemonic REPS). The trial was scored "2" if the same response was observed, and "1" otherwise. On replication trials in which the same pair was heard, but in inverted order (REPD), "2" was recorded when the responses differed, and "1" otherwise. Thus, "2" indicated consistency, with the separate tallies being recorded for subcategorizations of the reliability checking.

The previous paragraphs describe the translation into analysis code of the responses to specific stimulus pairs. This stage produced representations of the data as if the stimuli had been administered in a standard rather than in a random order. There is, however, another type of basic information retrievable at this point

in the analysis, and it revolves about the judgmental loop between the PEPO, PERE, and PORE stimulus pairs. Each of the three stimuli is a member of two of the pairs, and each is compared with the other two. This exhaustive comparison lends itself to two different types of mathematical procedures: Thurstonian scaling and transitivity analysis.

Thurstonian scaling (Torgerson, 1958) is a very simple extension of the present data, and we can take advantage of its operations to place the original stimuli on an interval scale. If POST is preferred to PRE by a certain proportion, then a specifiable psychological distance between those two utterances is implied. The existence of a three-part network of judgments of this triad allows the position of all three stimuli to be determined by responses to two stimulus pairs. Since the resultant values for PRE, POST, and RETENTION are on an interval scale, they will have to be transformed to a common baseline for comparison purposes. This will be further discussed below, when the Thurstonian analysis itself is presented. The concept was introduced here to facilitate the explanation of transitivity.

Since the three stimuli are compared exhaustively, we are provided with a second avenue for checking the reliability of the J's responses. We are further given the opportunity for a basic check on the validity of the main effect of the experiment. Since each J states his preferences for three pairs, the resultant set of three decisions can take on eight possible configurations. Of the eight, six reflect perceptual orderings which are internally consistent, and two produce nonsensical orderings of the original three stimuli. For example, the triad "2,2,2" as judgments of PEPO, PERE, and PORE pairs is equivalent to the J's statement "POST is better than PRE, RETENTION is better than PRE, and RETENTION is better than POST." This is also equivalent to the placement of the three stimuli in the order PRE, POST, and RETENTION on a preference continuum. The triad "2,2,1" may be translated into a statement similar to the above, save that the final preference is "POST is better than RETENTION." It is equivalent to the ordering PRE, RETENTION, and POST as most preferred. There are four other triads which can be translated into unambiguous preference orderings, and which are therefore transitive. The triads "1,2,1" and "2,1,2" cannot. The translation of the first of these is "PRE is better than POST; RETENTION is better than PRE; and POST is better than RETENTION."

The existence of a transitive relationship in the judgment triad is evidence that should increase our faith in the behavior of the J. However, it should be remembered that even if the J responds randomly, we would expect to obtain transitive triads an average of 6/8, or 75%, of the time. Therefore, the actual value for the number of triads producing any transitive ordering (mnemonic TRNA) must be compared with the expected value (75% of the total triads from which the sample is based) for evaluation of the strength of the effort. Significant increases in TRNA should strengthen our confidence in the consistency of the Js. The input data for those eventual reliability statistics were computed at the basic processing stage, where each J's response triad for each of the 9x24 words was scored positively for TRNA if it had any value other than "1,2,1" or "2,1,2."

Further consideration of the meanings of the triads "2,2,2" and "2,2,1" will reveal that not only do they bespeak a transitive, and hence reliable, relationship between the Js' responses, but such response triads also provide a bit of positive evidence toward the conclusion that there is a valid training effect to be observed in the stimuli themselves. For, if a given J places the PRE recording at the bottom of his preference order, this means that he has indicated to us that the treatment administered to S has improved his rendition of a particular word. Each triad was scored positively on the second transitivity criterion (TRNB) if it was either (2,2,2) or (2,2,1). The expected number of occurrences of TRNB is 2/8, or 25% by chance.

Now, of course, the two above triads do not have identical implications for the outcome of the training evaluation; their common ground is that PRE is judged poorest, and therefore training must have had some benefits. But what of the retention interval? A separate tally was made of the triads where (2,2,1) occurred; this was called TRNC, and indicates the judgment that RETENTION performance, while better than PRE, was surpassed by POST, i.e., that there was loss of performance caused by the no-training interval. The expected value of TRNC is 1/8, or 12.5%. The final transitivity tally, TRND, indicates the number of (2,2,2) triads, where S continues to improve through the retention interval. Its expected value is the same as in TRNC.

That completes the description of the basic processing of the data. To summarize: For each of the 24 critical words spoken by the Ss, several numbers had been generated from the responses of each J: PFPO, PLLO, PLRT, and PORL, standardized judgments

of specific pairs of utterances; REPS and REPD, simple measures of consistency; and TRNA, TRNB, TRNC, and TRND, measures of consistency and validity. This was the data format for the balance of the analysis. At no time were the data of a particular J used to conclude anything about the behavior of an S. The group of Js was used as a homogeneous panel, polled for its opinion on a large number of word pairs. The results for any word pair were expressed simply as a number ranging from 0 to 32. Magnitude of the difference between the members of the pair was considered to be a monotonic function of the output number, according to the standard models of psychological scaling and distance specification.

The time is now appropriate for an explicit statement of the logical paradigm for this analysis experiment. The Js were asked to state their preferences on a large number of pairs of words spoken by the Ss at different points in time. Their task might be viewed as one of simple psychological scaling, in which the amount of accent change reflects itself in the amount of agreement among the Js for a given word pair. If the Js consistently pick POST over PRE, this indicates that the stimuli actually differed in the direction predicted by training. The various psychological distances separating the four measurement points in time can be simply derived from the data. Remember that the words themselves are not the stimuli; rather, the stimulus is the time intervening between sampled utterances of the given word. We naturally expect all Ss to manifest some effect of training, but what of the treatment effect? It should be possible to demonstrate that the experimental treatment produces more improvement than the control treatment.

Overall analyses of variance. As a first attempt to speak to this issue statistically, several two-way mixed analyses of variance were performed. The format for the first set was a 9-column (Ss) by 24-row (individual critical words) matrix. Within a set,

nine analyses of variance were run, the only change from the above ten data tallies being that REPS and REPD had to be collapsed onto a single analysis for technical reasons. The generation of the PEPO analysis of variance matrix will be described. The maximum value in any cell was 32, which would have been obtained if all 32 Js had responded with a preference for the POST utterance for a given word as spoken by a particular S. To provide this matrix, the analysis software referred to the data structure, within which the responses of each J as processed by the first stage were organized in terms of word and S, and each S-word comparison was summed across all Js, producing the 216 cells. The three F ratios available from the analysis are used to evaluate treatment effects, word effects, and interaction between the two.

The first nine analyses of variance produced no consistent pattern of significance. In no case did a strong treatment effect emerge, with the great bulk of the treatment F-ratios falling in the region from .1 to .2. There was even less consistency among the word and interaction F-ratios. The existence of about only as many significant F-ratios as might be expected by chance in the 27 computed in the nine analyses of variance does not bespeak reliable effects, and so a table of their values is not provided. The small size of the treatment F-ratios implies the existence of a significant amount of variance in the data, which may well have blocked the extraction of any differential effect of treatment.

The major source of the variability of the data was the fact that the stimuli being judged—the utterances of the Ss—were highly variable themselves; and that this source, interacting with the normal response variability of the Js, had been amplified beyond the capability of the statistics. Some method was needed to reduce the stimulus variability, so that the postulated effects might have a chance of visibility.

Specialized statistics. To achieve this goal a large number of statistics was derived from the data base. To effect the maximal possible amount of data consolidation we collected the data from all control Ss and all experimental Ss into two separate groups, and, further, collapsed the responses of the Js to all 24 words within those two groups. This produced 12 pairs of numbers with one member of each pair summarizing the data of all words spoken by all experimental Ss. These 12 pairs are divisible into three types: actual pair preference, transitivity, and response reliability. The first two of these types will be presented.

Table 1 is divided into four sections, each of which has four columns. Each of these columns contains data derived from one of the four basic pairs of comparisons between the recording sessions. The four sections of Table 1 present alternative modes of inspection of the same basic data.

Table 1A gives the percentage of the time that the 32 Js responded with preference for the second members of the pairs indicated in the column headings. Since the number of experimental Ss exceeded by one the number of control Ss, there is a difference in the actual numbers of responses upon which the percentages are based. For the experimental Ss, each percentage summarizes 5 (Ss) x 24 (words) x 32 (Js' responses), or 3840 responses; the controls produced 4x24x32, or 3072. Of the 3840 opportunities that the Js had to respond to a PRE-POST comparison of words spoken by experimental Ss, they chose the POST version 62% of the time. Corresponding figures for other comparisons and for the data for the control Ss are similarly derived. The percentage differences

Table 1

Group Data - Pair Preferences

	PRE-POST	PRE-LOS	PRE-RET'N	POST-RET'N
A.	Percentage of Responses on Second Member of Pair			
Expected Value	50	50	50	50
Experimental (N=5)	62.1	59.0	62.6	50.9
Control (N=4)	61.0	58.3	61.4	49.9
E-C	1.1	0.7	1.1	1.0

	PRE-POST	PRE-EOS	PRE-RET'N	POST-RET'N
B.	Average Number of Second-Member Preferences per S			
Expected Value	384	384	384	384
Experimental (N=5)	476	453	480	396
Control (N=4)	468	447	471	383
E-C	8	6	9	7

	PRE-POST	PRE-LOS	PRE-RET'N	POST-RET'N
C.	Standard Score of Observed Deviation from Chance			
Expected Value	0	0	0	0
Experimental (N=5)	14.9	11.7	15.6	1.1
Control (N=4)	12.2	9.2	11.7	-0.1
E-C	2.8	2.0	3.9	1.2

	PRE-POST	PRE-EOS	PRE-RET'N	POST-RET'N
D.	Two-Tailed Probability of Occurrence of Mean by Chance (* Means <.001)			
Experimental (N=5)	*	*	*	(illegible)
Control (N=4)	*	*	*	(illegible)
E-C				

given in the fourth row of Table 1A are the representation of the treatment effect, while the disparities among the values expected according to chance (given in the first line) and the actual percentages are the representation of the effects of time, or the training effect. It is impossible to evaluate either effect meaningfully at the level of percentages or percentage differences, since there is no measure of the variability that might be expected according to the null hypothesis. It is, of course, obvious that no strong difference exists between the POST and RETENTION utterances in this analysis, since the percentages are so close to chance levels. It is also clear that the EOS utterances are less discriminable from the PRL than are either the POST or RETENTION utterances, for both experimental and control Ss.

Table 1B gives the analogous numbers of responses obtained from the Js, on a per-subject basis to cancel group size differences. Each S's 24 words produced 24x32, or 768 responses; by chance, 384 of these could be expected in either category. The actual averages numbers of responses for the first three pairs are radically different from chance expectations, lending strong support to training effects.

This effect is specifically evaluated in Table 2C, by reference to the expected value and theoretical standard deviation of the binomial distribution. The actual numbers of responses on the second members of the pairs were converted to standard scores by the conventional formula:

$$\text{standard score} = \frac{\text{(actual value - expected value)}}{\text{((total N)(probability of response A)(probability of response B))}^{1/2}}$$

where the numerator is the deviation from expectation and the denominator is the standard deviation of the sampling distribution of the binomial distribution with (in this instance) equiprobable alternatives under the null hypothesis. In the case of the PRL-POST comparison from the experimental Ss, the actual number of POST preferences was 2383 of a total N of 3840, with an expected value of 1536 for chance performance. The standard deviation of the binomial is 30.984, producing a standard score of 14.9, given in the second row of Table 1C. Chance performance would have been indicated by a score of 0 in this transformation.

Since there was no directionality implied in the derivation of the standard scores and since there was no a priori expectation

that a training effect in the direction of improvement would result from the treatment of the Ss, the correct procedure for evaluating the significance levels of the standard scores in Table 1C was a two-tailed test. Table 1D contains the output of a subroutine which, when given as input a standard score, computes the area under the normal curve lying distal to the absolute value of the input, i.e., gives the probability that a standard score value, as extreme or more extreme than the input, would occur under the null hypothesis that the sample was taken from a population of mean zero and standard deviation 1.0. The asterisks indicate that there is a strong effect of training in both experimental and control groups. The only results that are highly probable under the hypothesis of no discriminability between the members of pairs, arise in the POST-RETENTION comparisons. The treatment effect, if it is there at all, will have to be extracted differentially from the very strong training effect.

The final line of Table 1C gives a first look at the treatment effect per se. In the same terms as the training effect—standard scores—we see that the experimental Ss are consistently stronger than the control Ss in their improvement through time. Even in the retention interval, their performance does not decline as does that of the controls; and, while the significance level of the difference between the standard scores cannot be evaluated due to lack of a well-defined sampling distribution, the consistency of this treatment effect is encouraging.

Before we leave Table 1 for consideration of the transitivity data, a final note on the EOS "warm-up effect" is appropriate. It can be seen from the relative magnitude of the PRE-POST and PRE-EOS comparisons for both groups, that, had direct comparisons between POST and EOS been made, POST would have been preferred. Thus, our informal notion that Ss improved within a session is not borne out by comparison with the effects of the full training procedure. It might seem that the most direct comparison to speak to this issue might have been the POST-EOS pairs themselves,

but the PRE-EOS pair was used to maximize its commonality with the PRE-POST and PRE-RETENTION pairs.

Table 2 presents the second group of four summary statistics obtained from the collapse within treatment groups and across words. As described above, the four transitivity criteria were tallied for each triad of judgments produced by a given J to the various utterances of each S. For the five experimental Ss, the total number of response triads considered was 5x24x32, or again 3840.

It is clear in Table 2A that all four transitivity tallies have recorded more positive instances than might have been expected by chance. TRNA, which reflects the existence of internally consistent responses to the triads, regardless of the resultant ordering, has no specific bearing on the treatment or training effect; its size is most relevant to the reliability of the Js' responses. One of the major factors holding down the value of TRNA in this situation might well have been the multidimensionality of the stimuli being judged. If different errors are made in the three members of a triad, or if J focuses on one sub-aspect of vowel pronunciation, such as duration, instead of another, such as quality, at different pairs within the triad, internal consistency may be lost. The fact that this did not happen is of interest, though the simple percentage increase over the expected value of 75% cannot in itself be evaluated without reference to the standard scores presented in Table 2C. The same strictures apply to the percentage scores for TRNB, C, and D, though they seem even further removed from their expected percentage values. We note further that the percentage differences between experimental and control Ss are in the proper direction for a positive-treatment effect, and we see the reflection of this effect in the averaged data for the two groups of Ss in Table 2B, but it is Table 2C where the treatment effect is most graphically displayed.

The standard scores displayed in Table 2C were derived from the same formula as that given above, but the denominators of the computations reflected the asymmetric probabilities for the occurrence of the various criteria. The TRNL criterion, which was

Table 2

Group Data - Transitivities

A. Percentage of Judgment Triads Producing A Criterion Response

	TRNA	TRNB	TRNC	TRND
Expected Value	75	25	12.5	12.5
Experimental (N=5)	83.7	42.5	21.2	21.3
Control (N=4)	83.2	41.0	20.7	20.3
E-C	0.5	1.6	0.6	1.0

B. Average Number of Triads Meeting Criterion per S

	TRNA	TRNB	TRNC	TRND
Expected Value	576	192	96	96
Experimental (N=5)	643	326	163	163
Control (N=4)	639	314	158	155
E-C	4	13	5	8

C. Standard Score of Observed Deviation from Chance

Expected Value	0	0	0	0
Experimental (N=5)	12.5	25.1	16.4	16.4
Control (N=4)	10.5	20.4	13.7	13.0
E-C	2.0	4.7	2.7	3.4

D. Two-Tailed Probability of Occurrence of Data by Chance (*Means $< .5 \times 10^{-5}$)

Expected Value				
Experimental (N=5)	*	*	*	*
Control (N=4)	*	*	*	*
E-C				

satisfied by a judgmental triad that is not only internally consistent but places the PRE utterance at the bottom of the implied preference continuum, is oversubscribed for both groups of Ss. The control Ss manifested a very strong effect of training according to this measure, being 20.4 standard deviations above what might have resulted from chance; but the experimental Ss were a full 4.7 standard deviations more removed from the mean. Though the absolute difference between the two treatments is small, it looms large when placed in the perspective of statistical expectation.

The constituents of TRIL are TRIC and TRID, with the latter recording the instances in which a judgmental triad implying the preference ordering PRE, POST, and RETENTION. Positive TRIC instances reflect consistent triads showing overall improvement with respect to the PRE utterances, but also showing decrement over the retention interval. The differences between the standard scores of the two treatment groups for these criteria remain high, and the further fact that the treatment difference is larger in TRID than in TRIC (3.4 versus 2.7) is further evidence in favor of the efficacy of the full feedback capabilities of the API system in improving and maintaining the accents of the Ss in the experimental group.

It should be remembered that the extraction of the above-noted treatment effects was achieved only by means of a large amount of data consolidation. In performing this consolidation, information about specific words and Ss has been sacrificed in order that the treatment effect might be teased out of the immensely stronger training effect. Table 1 was derived from the data base after a collapse across words and Ss, within specific stimulus pairings. The transitivity statistics possess the further characteristic that triads of judgments are handled together. This approach provided the most global inspection of the data, and, consequently, it encountered the most success.

Let us consider one further treatment of the data. It has been stated above that in one sense the point of the analysis is to scale the effect of training on the utterances of Ss. This was expressed by a score of "correct" when any J responded in such a way as to place two stimulus utterances in "proper" temporal order. The outcomes of many such judgments across various Js, Ss, and words may thus be used as a metric for the psychological distances between the accent levels of the Ss as recorded at different times during the experiment. The standard scores of the four comparison pairs presented in Table 1C are such distances. Three of them may be combined explicitly by Thurstonian scaling algorithms since they form an exhaustive set of judgments for three stimuli, and Table 3 summarizes the outcome.

Table 3

Group Data
Scale Values with Pre Set to Zero

	POST	RETENTION
Experimental (N=5)	0.304	0.324
Control (N=4)	0.283	0.286
E-C	0.021	0.038

The psychological distances between the perceived accents on the three critical days were computed separately for the two treatment groups. This yielded positions for the three test days on an interval scale with no inherent origin. Comparisons between the groups were made possible due to the fact that the pretest utterances had all been rated before the start of the experiment, and placement within the respective treatment groups had been done in a pairwise manner, which matched the apparent accents between the two groups, in addition to matching their exposure to English. This prematching justified the addition of a number to each of the two interval scales such that the pretest scale value was changed to zero. This was done by adding to each triad of scale values the pretest scale value of the triad, and it had the effect of transforming interval data into ratio scale values, based on a common zero point set at pretesting time. The units of the scales are arbitrary. Their directionality is a restatement of the strong training effect, since both groups produce positive values for POST and RETENTION. The experimental group is again seen to be stronger at the conclusion of training, and to retain and even slightly consolidate that strength during the retention interval while the control Ss remain essentially constant.

This psychological scaling completes the global analysis of the data. It has produced what is hoped is a coherent picture of the overall effects of the experiment. The structure of the data has been shown to be internally consistent by the interdependent statistics employed to probe it.

APPENDIX 4.3: REDUCED-VOWEL TONGUE POSITION DISPLAY

We assume that a set of English words has been selected with the properties that they contain at least one unstressed syllable, and syllables are separated by either stop gaps or voiceless consonants, e.g., "multiply," "about," "photograph," etc.

Syllables are identified from the filter-bank input by an algorithm that is illustrated below. A time function, F(nT), the sum of filter outputs 2 and 3, is chosen to emphasize the low-frequency energy which is characteristic of voicing in vowels. Significant peaks in this function indicate the approximate midpoints of syllables. For a peak to be called significant, it must have the property that adjacent peaks are separated by valleys of at least 15 dB less than the magnitude of the peak. This is determined by the following two-state algorithm which starts at time nT=0 in state 1 with LOCMIN = 0, WCMAX = 0, and MSYC = 1.

```
STATE 1.0  n=n+1
           if F(nt)>LOCMAC       go to 1.1
           if F(nt)<LOCMAC-40    go to 1.2
           go to 1.0
      1.1  LOCMAX = F(nt)
           TMAX = n
           go to 1.0
      1.2  LOCMAX = 0
           LOCMIN = F(nt)
           TSYL(NSYC) = TMAX
           NSYL = NSYL+1
           go to 2.0
STATE 2.0  n=n+1
           if F(nt)<LOCMIN       go to 2.1
           if F(nt)>LOCMIN + 40  go to 2.2
           go to 2.0
      2.1  LOCMIN = F(nt)
           go to 2.0
      2.2  LOCMIN = 1000
           LOCMAX = F(nt)
           TMAX = n
           go to 1.0
```

When n>NMAX, then the time of the desired syllable, m, equals TSYL (m).

Once the relevant syllable has been isolated, the analysis of the vowel nucleus proceeds in a way that is virtually identical to the method of handling stressed vowels. A complete description of the vowel display algorithm for single syllable words can be found on pages 52 thru 57 of Semiannual Technical Report No. 7 (Kalikow and Klatt, 1970). One minor difference

between the previous vowel display algorithm and the reduced vowel display is that, in the reduced vowel display, the entire vowel nucleus trajectory is displayed instead of attempting to suppress sample points at the beginning and end of the trajectory to reduce consonantal influences.

As in the original vowel display, the criterion rectangle appearing on the oscilloscope screen has a size and position that depends on each individual word. Typically, the rectangle is somewhat larger than for stressed vowels and remains near the center of the vowel triangle.

APPENDIX 4.4: EXPERIMENT II DATA ANALYSIS

Table 4 gives data for the four individual Ss and for their average performance within the two displays. Each S's performance within a given display is summarized by four numbers. For example, S1's 12 practiced words in the AVOD were collected into 12 PRE-POST pairs and resulted in 120 judgments (across words and Js). Of these, 104 were "correct". Subject 1's overall performance was therefore 152 "correct" out of a possible 180 for a total percentage of 84.4. Such an occurrence is highly unlikely under the null hypothesis of no training effect, as indicated by the standard score of 9.2. Across the four Ss and two displays, training effects were observed in all but one instance.

An investigation of the generalization effect was not carried to the level of individual Ss, since one always runs the risk of chance significance when the number of tests proliferates. The final two columns of Table 4 give the averaged data for the four Ss in the two displays. These were discussed in Section 3.2.4 above.

The sources of these data are made clearer in Tables 5 and 6. These contain judgment data by S and word, and certain summarization statistics. The specific words trained and tested are shown, though not in the actual orders used. They are grouped in terms of type or location of distinctive features: by kind of initial consonant for the AVOD words (/t/, /p/, or /k/), and by location of the syllable containing the reduced vowel for the RVTPD.

Inspection of the pattern of response across the four Ss indicates that there is by no means a high rank-order correlation between performances on the two displays, indicating that either Ss possessed different starting levels of accent in the two areas or that they improved differentially. For example, while S3 improved almost totally in the AVOD, she did not show the greatest improvement for the RVTPD.

Table 4

Percentages of Post Preferences
by Subject and Word Group in Two Displays
Ten Judges

AVOD

	S1	S2	S3	S4	Average	Std. Score
Practiced	86.6	69.1	97.5	89.1	85.6	15.7*
New	80.0	76.6	96.6	93.3	86.6	11.4*
Total	84.4	71.6	97.2	90.5	85.9	19.3*
Std. Score	9.2*	5.8*	12.7*	10.9*		

RVTPD

Practiced	48.3	60.0	60.8	88.3	64.4	6.3*
New	51.6	53.3	60.0	58.3	55.8	1.8†
Total	49.4	57.7	60.5	78.3	61.5	6.2*
Std. Score	-.15 ns	2.1*	2.8*	7.6*		

*: significant at or beyond .05 level, 2-tailed

†: significant at .07 level, 2-tailed

ns: not significant

The leftmost columns of Tables 5 and 6 give the strengths of the displays by type of stimulus word. The new words, while showing a significant generalization effect in the AVOD, are still preceived as improving less than trained words; and the pattern of improvement differs between the two groups of words. Interestingly, the generalization effect seems strongest in new reduced-vowel words where the reduced vowel is contained in the first syllable. There seems to be no reliable generalization of more complex new reduced-vowel words. Due to the relatively constricted amount of data, such detailed analyses as presented here may tend to overemphasize the natural variability inherent in accent-judgment data, and the reader should keep this in mind when inspecting specific response categories. The overall analysis of Table 4, relying as it does on the maximally inclusive statistical tests, must remain the touchstone. The subsequent more detailed analyses do not affect its conclusions as to the reliability of the training effects for both displays.

Table 5

Number of Judges Preferring Post Utterances
by Subject and Word - AVOD Display

Ten Judges

INITIAL CONSONANT	PRACTICED WORDS	S1	S2	S3	S4	SUM	RATIO	PERCENT	STD SCORE
T	1. team	8	6	10	8	32			
	2. tune	9	10	10	9	38	141/160	88.1	9.7*
	3. turn	10	6	9	10	35			
	4. tot	7	9	10	10	36			
P	5. poor	10	6	10	9	35			
	6. pond	10	6	10	8	34	141/160	88.1	9.7*
	7. pearl	9	8	10	8	35			
	8. peel	10	7	10	10	37			
K	9. curt	9	7	10	10	36			
	10. key	6	10	9	8	33	129/160	80.6	7.8*
	11. cool	10	2	10	9	31			
	12. cot	6	6	9	8	29			
		104	83	117	107	411			
	NEW WORDS								
T	1. tell	9	7	10	10	36	67/80	83.8	6.0*
	2. took	6	6	10	9	31			
P	3. pen	9	7	9	10	35	72/80	90.0	7.2*
	4. poll	10	8	9	10	37			
K	5. kit	10	9	10	8	37	69/80	86.3	6.5*
	6. cow	4	9	10	9	32			
		48	46	58	56	208			

*: Significant at or beyond .05 level, 2-tailed

Table 6

Number of Judges Preferring Post Utterance
by Subject and Word - RVTPD Display

Ten Judges

TARGET SYLLABLE	PRACTICED WORDS	S1	S2	S3	S4	SUM	RATIO	PERCENT	STD. SCORE
First	1. (su)ppose	8	10	9	10	37			
	2. (di)stinguish	5	6	5	8	24	$\frac{105}{160}$	65.6	4.0*
	3. (spe)cific	6	5	8	8	27			
	4. (de)fend	4	1	9	3	17			
Second	5. tes(ti)fy	8	8	9	10	35			
	6. diff(i)cult	4	7	3	9	23			
	7. fea(si)ble	6	4	4	10	24			
	8. ex(e)cute	4	4	1	10	19	$\frac{204}{320}$	63.7	4.9*
	9. of(fi)cer	9	10	10	10	39			
	10. opp(o)site	1	6	8	10	25			
	11. suff(o)cate	2	6	4	8	20			
	12. sen(si)tive	1	5	3	10	19			
		58	72	73	106	309			
	NEW WORDS								
First	1. (de)tect	6	3	5	7	21	$\frac{48}{80}$	60.0	1.8†
	2. (de)scription	6	9	8	4	27			
Second or Third	3. iden(ti)fy	8	7	2	7	24			
	4. mu(si)cal	5	3	7	6	21	$\frac{86}{160}$	53.7	.95 ns
	5. tel(e)phone	2	6	9	4	21			
	6. mic(ro)phone	4	4	5	7	20			
		31	32	36	35	134			

*: Significant at or beyond .05 level, 2-tailed
†: Significant at or beyond .07 level, 2-tailed
ns: non-significant

5. REFERENCES

Lado, R. Language testing. McGraw-Hill, New York, 1961.

Lindquist, E.F. Design and analysis of experiments in psychology and education. Houghton Mifflin, Boston, 1953.

Torgerson, W.S. Theory and methods of scaling. Wiley and Sons, New York, 1958.

Winer, B.J. Statistical principles in experimental design. McGraw-Hill, New York, 1962.