DOCUMENT RESUME

ED 100 353 16 IR 001 451

AUTHOR	Nickerson, Raymond S.; And Others
TITLE	A Computer-Based System of Speech-training Aids for the Deaf: A Progress Report.
INSTITUTION	Bolt, Beranek and Newman, Inc., Cambridge, Mass.
SPONS AGENCY	Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C. Media Services and Captioned Films Branch.
REPORT NO	BBN-2901
PUB DATE	14 Sep 74
CONTRACT	OEC-0-71-4670(615)
NOTE	57p.; For related documents see IR 001 449 through 455
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	*Computer Assisted Instruction; Computer Graphics; *Deaf Children; Deaf Education; Deaf Research; Educational Games; Electromechanical Aids; Handicap Detection; Speech Education; Speech Evaluation; Speech Improvement; *Speech Therapy
IDENTIFIERS	Nasality

ABSTRACT

Progress on an effort to develop a computer-based system of speech-training aids for the deaf is reported. The systems is described, as are four different types of visual displays. One type, showing speech parameters as functions of time, is discussed in detail. The problem of assessing speech performance and the possibility of giving the computer some evaluation capabilities are briefly considered. (SK)

BOLT BERANEK AND NEWMAN INC

CONSULTING · DEVELOPMENT · RESEARCH

Report No. 2901
Job No. 11580

A COMPUTER-BASED SYSTEM OF SPEECH-TRAINING AIDS FOR THE DEAF: A PROGRESS REPORT

14 September 1974

Submitted to:

Media Services and Captioned Films Branch
Division of Educational Services
Bureau of Education for the Handicapped
U.S. Department of Health, Education, and Welfare
Washington, D. C. 20202

Table of Contents

Abstract

Progress on an effort to develop a computer-based system of speech-training aids for the deaf is reported. The system is described, as are four different types of visual displays that have been programmed to date. One type of display, which shows various speech parameters as functions of time, is discussed in detail. The problem of assessing speech performance and the possibility of giving the computer some evaluation capabilities are briefly considered. The importance of close collaboration between researchers and teachers on efforts to develop innovative training aids is emphasized, as is the need to resolve some basic pedagogical issues.

A Computer-Based System of Speech-Training Aids for the Deaf: A Progress Report

R. S. Nickerson, D. N. Kalikow, and K. N. Stevens

This paper is a progress report on an effort to develop a computer-based system of speech-training aids for the deaf. The project was begun with the assumption that an attempt to *design* such a system would probably fail, and that a more promising approach would be to attempt to evolve one through use. Accordingly, a system incorporating some of the capabilities that it was thought would be useful for speech training was developed, and installed at the Clarke School for the Deaf where it is now being used on an experimental basis in a remedial speech-training program. The expectation was that the capabilities of the system would be modified and extended as attempts to use it provided insights concerning what features it should have. To ensure that such insights do in fact guide the system's evolution, developers and users are engaged in a continuing dialogue concerning the desirability and feasibility of specific modifications and extensions, both in the training procedures that are used in conjunction with the system and in the characteristics of the system itself. The purpose of this paper is to describe how certain aspects of the system have evolved over the first several months of use.

The general considerations that governed the initial development of the system were the following. Deaf students receive only minimal acoustic information from the speech of others and from their own vocalizations. The speech skills they acquire are based on cues they receive from their residual hearing and from visual observations of the gestures of others. Often these skills are inadequate or incorrect, and the students thus need special training in order to help them to produce intelligible speech. As a part of this training, it is customary for a teacher to produce speech-like patterns or to describe the patterns to the child and for the student to try to imitate these patterns. The student is encouraged by the teacher if he produces the correct speech gesture. Three problems arise in this kind of training situation: (1) the relevant attributes of the speech sample produced by the teacher often cannot be seen, felt, nor heard by the student; (2) the student must rely on the teacher to indicate whether or not his production is acceptable; and (3) the teacher must make a subjective judgment as to the adequacy of the student's production. All three of these problems provide motivation for developing a set of displays for use in a speech-training situation.

In this paper we will not address the issue of how speech training might be most effectively carried out, if a system of the type described here were available. Many of the questions that need to be answered concern speech-training strategy in general and arise quite independently of the existence of any particular training aids. These questions include the following. In what order should various types of speech skills be developed? Should primary emphasis be placed on temporal and prosodic aspects of speech or on articulatory skills in isolated syllables? Should vocal exercises of various kinds be used before working on the production of speech material? What strategies will optimize the generalization of speech skills acquired in one phonetic context to other contexts? To the teacher of the deaf, questions such as these are, of course, of central concern. The primary aim of this project is not to answer these questions, but to develop a versatile tool that may be applied by a teacher to certain aspects of speech training regardless of what the answers are. The introduction of new tools often has the effect, however, of prompting changes in technique. It would be surprising if the existence of a system that facilitates the objective assessment of speech patterns and the evaluation of progress and failures of a student did not also provide some basis for modifications and improvements in speech-training methods.

The idea of using visual displays of speech parameters to aid in speech training of the deaf is, of course, very old. In recent years, numerous instruments have been developed to produce a variety of different visual patterns (Levitt, 1973; Pickett, 1968). The system described here incorporates within a single unit some of the kinds of displays described previously by others (although usually in modified form), as well as some new displays.

THE SYSTEM

The system is built around a small digital computer, the Digital Equipment Corporation PDP-8E. Speech information is obtained from a miniature accelerometer attached by thin double-stick tape either to the throat or the nose, and from a head-mounted voice microphone. The accelerometer (BBN Model 501), which is approximately .3 inches in height and diameter, and weighs about 1.8 grams, is used to simplify the extraction of certain parameters that are relatively difficult to derive from a microphone output. When the accelerometer is attached to the throat it gives a waveform that has periodic peaks at the frequency of the glottal output during voiced sounds. The output is fed to a pitch extractor circuit that measures the time between positive-going zero crossings of the waveform and reports the pitch periods to the computer. When attached to the nose, the accelerometer provides a signal that is a measure of the amount of acoustic coupling to the nasal cavity through the velopharyngeal port. In this case, the output, which is 10-15 dB higher when the velum is lowered--during nasalized sounds--than when it is raised, is fed to a component that rectifies and low-pass filters it and sends the result on to the computer. The use of the accelerometer for the acquisition of pitch and nasality information is described more fully by Stevens, Kalikow, and Willemain (1974). The output of the voice microphone is fed into a filter bank that reports to the computer the energy in each of 19 frequency bands within the range 100-6560 Hz. The parameters of these filters are shown in Table 1. Data from the pitch extractor or nasality circuit (only one of these components is operational at a given time in the current system) and the filter bank are sampled by the computer 100 times per second, and used to generate a variety of visual displays. Control inputs from the user are given to the computer via a set of push-buttons and analog knobs. For further details concerning the system, see Nickerson and Stevens (1972, 1973).

DISPLAYS

Displays are generated by four independent programs. We will refer to these programs as: (1) ballgame, (2) vertical spectrum, (3) cartoon face, and (4) time plot. We will describe the first three programs only very briefly, and concentrate on the fourth, inasmuch as this program has been used most extensively and has evolved to the greatest extent to date. The first two programs are also described in more detail in Nickerson and Stevens (1972, 1973).

Ballgame

The intent in developing the *ballgame* program was to implement a display that would be motivating to a young child, and

Table 1. Frequency parameters of Filter bank.

Filter	Lower	Higher	Center	Band
No.	Cut-Off	Cut-Off	Freq.	Width
	Hz.	Hz.	Hz.	Hz.
	cps	cps	cps	cps
1	100	440	260	360
2	260	620	440	360
3	440	800	620	360
4	620	980	800	360
5	800	1160	980	360
6	980	1340	1160	360
7	1160	1520	1340	360
8	1340	1700	1520	360
9	1520	1880	1700	360
10	1700	2060	1880	360
11	1880	2240	2060	360
12	2060	2420	2240	360
13	2240	2600	2420	360
14	2420	2780	2600	360
15	2600	2960	2780	360
16	2960	3560	3260	600
17	3560	4400	3980	840
18	4400	5480	4940	1080
19	5480	6560	6020	1080

that would permit him to develop some speech-related skills in a game-like situation. The game starts with several balls positioned in the lower left corner of the display. When the child is ready, he or the teacher presses a button which moves one of the balls to a little pedestal or takeoff point at the extreme lower left. When the child starts voicing, the ball begins to move at a fixed rate toward the "wall" at the right of the display, as shown in Figure 1.[1] The height of the moving ball is determined by the value of some parameter of the speaker's voice. To date, the only parameter that has been coded thi way is voice fundamental frequency (loosely called pitch); however, any other single-valued function could also be used. The child's task is to make the ball go through the hole in the wall. If he succeeds, the ball drops in the basket that is positioned to the right of the wall, and a smiling face appears in the upper right corner of the screen. If the ball is either too high or too low to go through the hole, it bounces from the wall back to the starting position, and the child may try again. Both the height of the hole in the wall and its width are adjustable by the teacher. Such adjustments are made by turning appropriate control knobs.

In a more complicated version of the ballgame display, a second wall may be added to the left of the first, its distance from the first wall also being adjustable with a control knob (see Figure 2). By placing the walls relatively close together and adjusting the holes to different heights, the teacher can define a task in

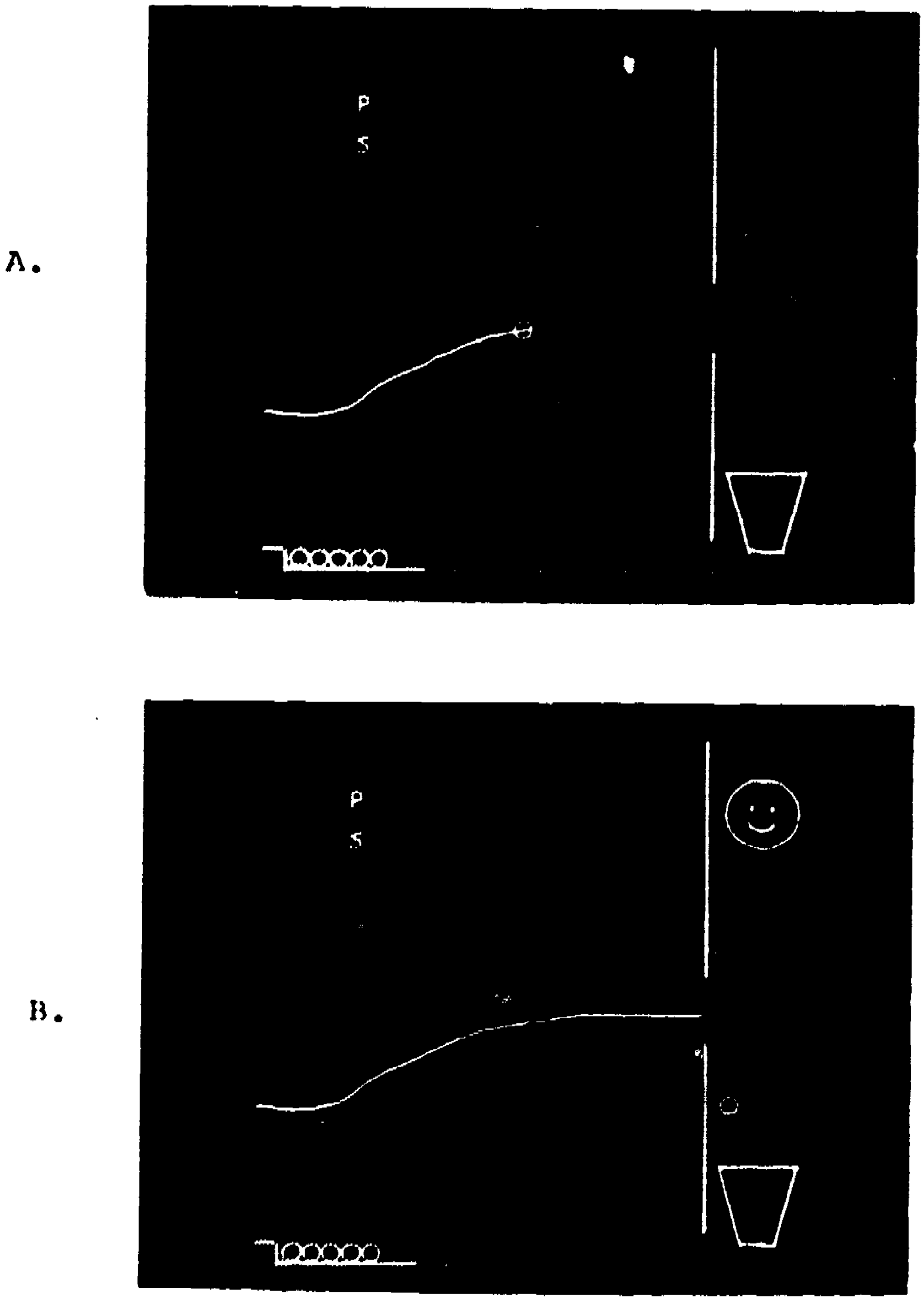


Fig. 1. Ballgame display. A: midway through an attempt; B: the attempt succeeds. See text.

which the child must make his pitch rise or fall by specific amounts within a certain time interval in order to be able to get the ball in the basket.

In one version of the display, not only is the height of the ball a function of pitch but the size of the ball is a function of speech intensity ("loudness"). In this case not only must the ball be at the correct height but it must be small enough to fit through the hole if a basket is to be scored. The positions of the holes and their widths (in Hz. and fractions of an octave, respectively) and the distance of each wall from the starting position (in milliseconds) are displayed on demand.

Vertical Spectrum

The operation of the vertical-spectrum display is illustrated in Figure 3. Each shape in this figure is determined by the frequency spectrum of the sound from which the shape is generated. Frequency goes from low to high, along the vertical axis, and the width of the shape at a given height is proportional to the energy (on a logarithmic scale) within a specific frequency band. Voiced and voiceless sounds are distinguished by the presence and absence of the horizontal lines, respectively. When the amplitude of the filter with the largest output exceeds a certain threshold value, the width of the pattern at this frequency no longer increases and all other widths are normalized to this maximum. Thus the shape and width of the pattern remain relatively independent of the voice level above this threshold.

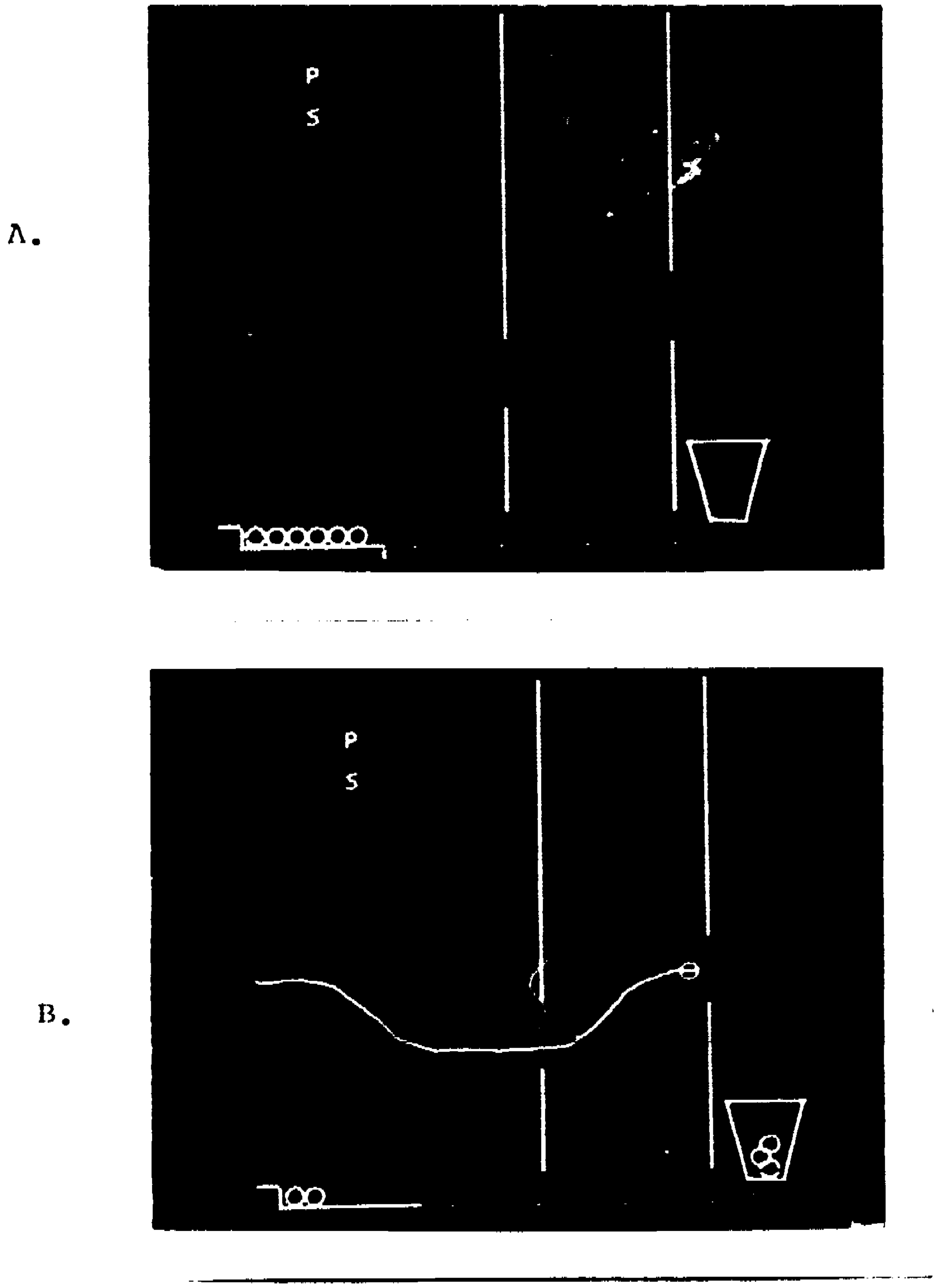


Fig. 2. Ballgame display, two-walled version. A: initial appearance. B: the speaker about to achieve his fourth success in the game.

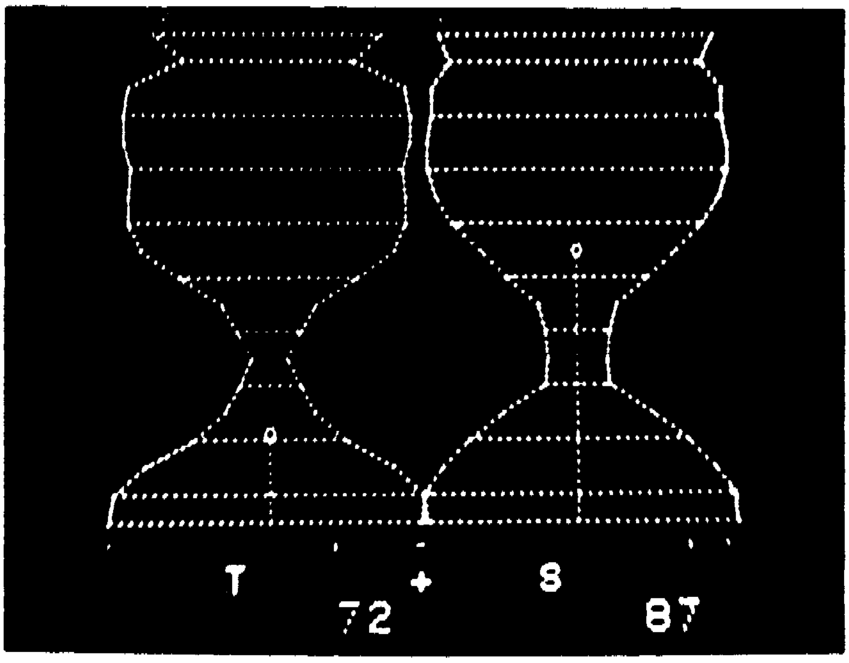


Fig. 3. Vertical spectrum display. Left and right figures produced by male and female speakers, respectively, during utterance of the vowel /i/ in the word "be."

The pitch of the speaker's voice is represented by the length of the "lollipop" rising from the center of the base. Note that in the example shown in Figure 3 the two shapes, both of which represent the vowel /i/, as in "be," are fairly similar in spite of the fact that the left one was produced by a male speaker with a fundamental frequency of about 100 Hz. and the right one by a female speaker with a fundamental frequency of about 210 Hz.

The right-hand shape in this display changes continuously as the student speaks, unless he is sustaining a steady sound, in which case the shape remains relatively constant. The shape on the left may be used as a standard that the student may be asked to match. In order to generate a standard, the teacher makes a sound and presses a button when he has produced the desired shape. When the button is pressed the display is "frozen," and the captured shape remains constant until it is explicitly changed. The student's (right-hand) display may also be frozen; and it may be transferred to the left side of the display where it can serve as a standard for future attempts at matching.

This display has a replay capability; whenever the display is frozen, the computer retains a record of the speech immediately preceding the instant represented by each shape on the display. The teacher can, through the use of the control buttons, instruct the computer to replay either stored speech sample and to redisplay it as it is being replayed. The replay can itself be frozen at any given point and inspected frame-by-frame, if that is desired.

Cartoon Face

The cartoon-face display is illustrated in Figure 4. The idea here was to develop a display in which several parameters of speech could be represented simultaneously in a single integrated scene. This display is the newest and least thoroughly tested of the four that have been developed. We expect it to change considerably as attempts are made to use it. At the present, four aspects of speech are represented by the features of the face: the presence of an "Adam's apple" on the throat signifies the presence of voicing; its height indicates the fundamental frequency of the voiced sound; loudness is represented by the size of the mouth; and the detection of an "s" or "z" is indicated by the appearance of the letter "s" or "z" in the cartoonist's balloon. A natural elaboration of this display would be the use of a blinking nose to indicate that the nasal energy has exceeded a preset criterion. But addition of this feature would require a modification to the system that would permit the measurement of pitch and nasality simultaneously.

The intent in representing speech parameters in this way was to present a scene that would be perceived by the viewer as a single thing, as opposed to a collection of independent symbols, and that would have some intrinsic appeal to a young child. It remains to be determined whether these objectives have been met, and whether the child can make effective use of the information that is encoded in this form.

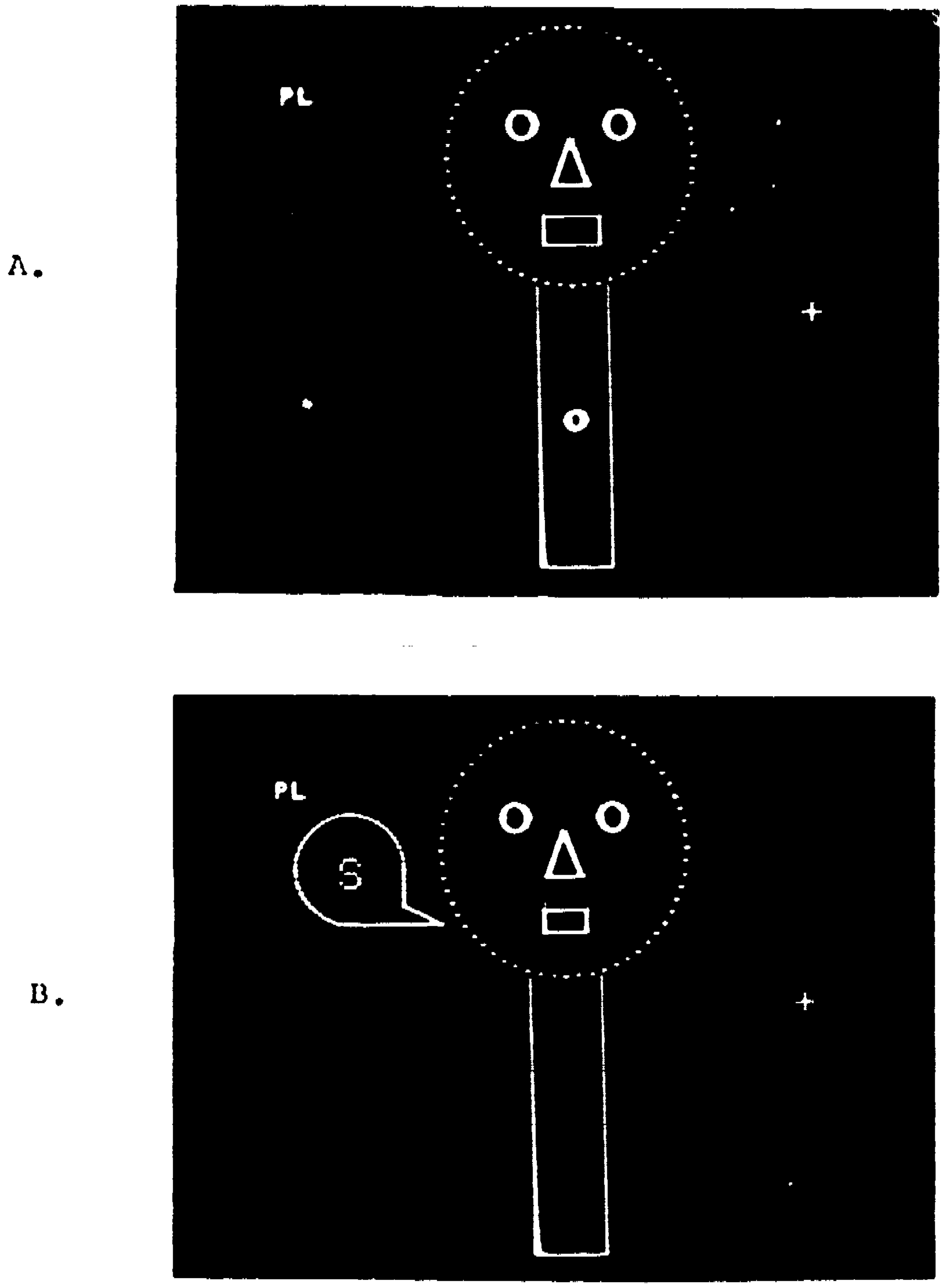


Fig. 4. Cartoon face display. A: responding to a voiced sample of speech, showing the "Adam's apple" at a position related to the most recently sampled pitch, and the mouth opened proportionally to the "loudness" of the voice signal. B: the display during an unvoiced sample meeting the criterion for the "s-detector."

Time Plot

The time-plot program has the capability for displaying severa? parameters of speech as functions of time. In the remainder of this paper, we will confine our attention to this program and discuss its capabilities and modes of operation in some detail.

As noted above, one of the ways in which the user gets control information to the computer is by means of a set of twelve push-buttons. The functions of the buttons are determined by the program that is operating in the machine. Associated with each of the four programs mentioned above there is an overlay for the button box that identifies the meanings of the buttons for that particular program. The push-buttons are made of translucent plastic, and each contains a miniature light bulb which is switched on or off by the computer; only buttons that represent viable options to the user are illuminated at any given moment. The significance of specific push-buttons will be discussed in the appropriate context below.

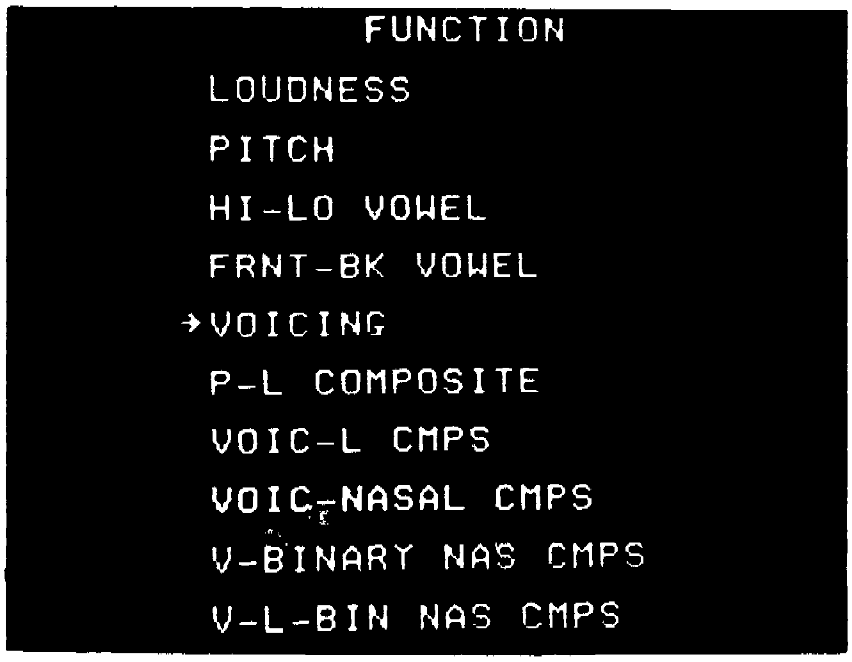
The options that are available to the user of the time-plot program are of two types: the function or functions to be displayed, and the mode of operation of the display. At any given time, some specific combination of the options in these two categories is in effect. To select or change an option, the user presses the "category" button on the button box. A press of this button causes a "menu" to appear on the display such as that shown in Figure 5. When a menu is displayed, an option is selected from it by means of the "option" button. Each time this button is pressed an arrow on the left of the display (pointing at "voicing" in Figure 5) is moved down one line. When the user has positioned the arrow beside the option he wants he presses the "return" button which invokes the selected display.

The menu shown in Figure 5 is shown for the sake of illustration only. It represents the functions that were implemented in the time-plot program at a particular stage of its evolution. It is not to be expected that all of these functions will be equally useful, or even that all of them will survive a period of testing. The list has already changed several times, and it is likely to undergo further changes as the program continues to evolve.

Function options. Figure 6a shows plots of speech amplitude, or what we loosely call "loudness," as a function of time for the utterance "What time is it?". This and all subsequent displays from the time-plot program are obtained by smoothing the original sampled data with a time constant of 25 milliseconds. In the top trace the emphasis was on "time," in the bottom one it was on "is." (Figures 6b and 6c will be discussed presently.) The algorithm that is used to calculate loudness is as follows:

$$\text{Loudness} = F_1 + F_2 - F_3 - F_4 + \log \sum_{i+1}^{16} A_i$$

where $F_i = \log A_i$, and A_i represents the rectified and smoothed output of the i^{th} filter of the filter bank. This is not, of course, a true estimate of loudness, but it is a measure that



```
          FUNCTION
     LOUDNESS
     PITCH
     HI-LO VOWEL
     FRNT-BK VOWEL
    →VOICING
     P-L COMPOSITE
     VOIC-L CMPS
     VOIC-NASAL CMPS
     V-BINARY NAS CMPS
     V-L-BIN NAS CMPS
```

Fig. 5. Time-plot program: FUNCTION menu.

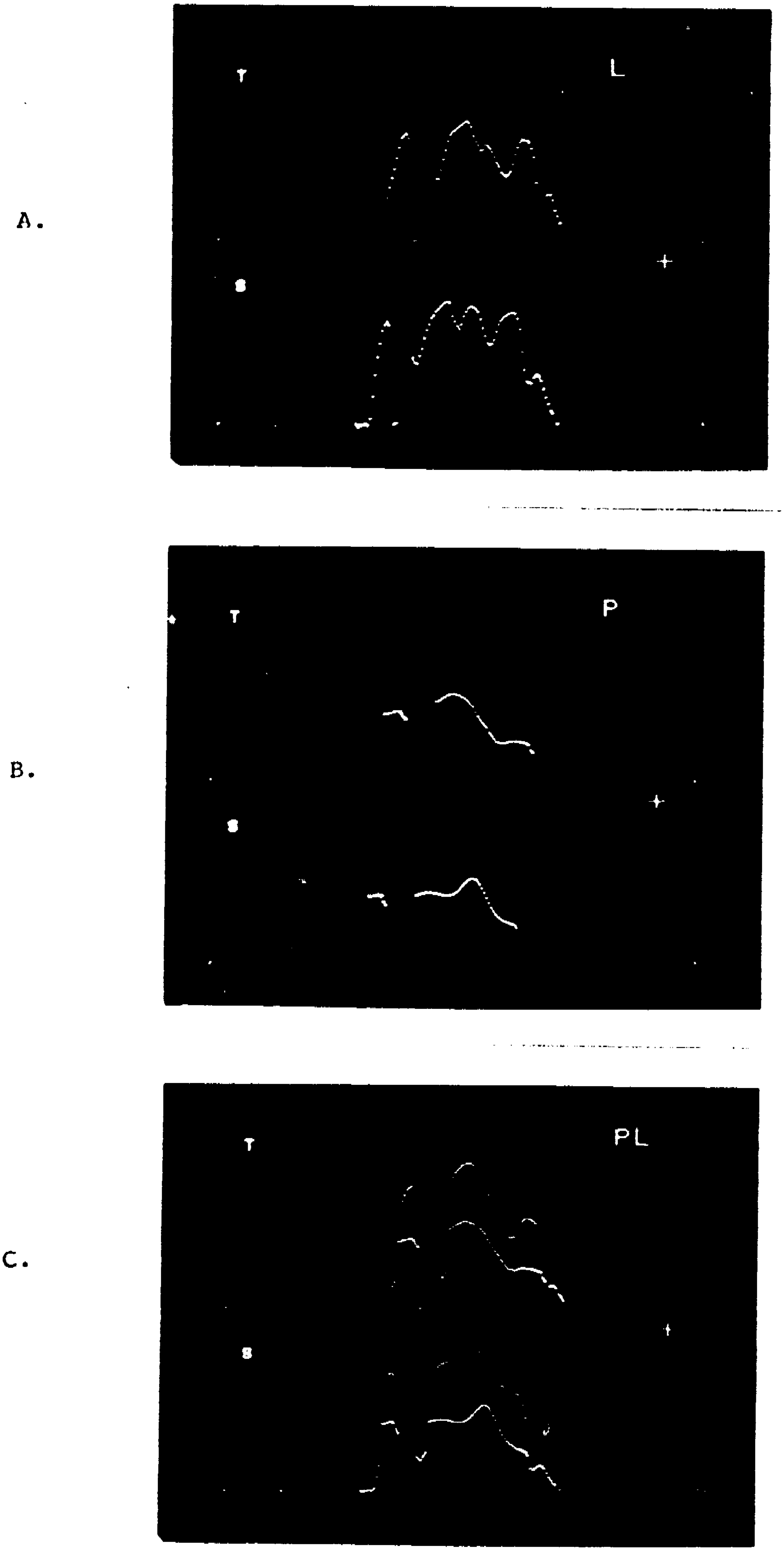


Fig. 6. Various displays of the data arising from two utterances of the sentence, "What time is it?" spoken by a male. In the top trace, the emphasis was on the word "time," and in the bottom one on the word "is." A: loudness; B: pitch; C: pitch-loudness composite.

brings certain intensity-related attributes of an utterance into evidence. The display tends to indicate the fluctuations in intensity from one syllable to the next with minima during the consonantal intervals and maxima during the vowels. The component $F_1 + F_2 - F_3 - F_4$ tends to correct for the inherent differences in amplitude between low and high vowels. It adds a correction for high vowels (with low first formant) and subtracts a correction for low vowels (with a first-formant peak in the vicinity of the fourth filter). This is intended to be only a partial correction, however, since it is desirable to show separate syllabic peaks in the function for an utterance like /ɑyɑ/. The display also shows the presence of weak voiceless consonants such as /f s š/ and plosive releases in syllable-final position, since it tends to emphasize high frequencies, where the filter bandwidths are greater. A composite display in which both voicing and loudness are represented (and which will be described below) is used more frequently than the simple loudness display. One application that has been made of such displays is the training of timing and rhythm.

The top trace in Figure 7 represents fundamental frequency as a function of time for the utterance "How are you?" with stress on "are"; the bottom trace is the same parameter for the utterance "I'm fine. How are you?", with stress in this case on "fine" and "you." Fundamental frequency is plotted on a logarithmic scale. Points are plotted only when the vocal cords are vibrating, that is, during voiced intervals of the speech. Segments of the display for

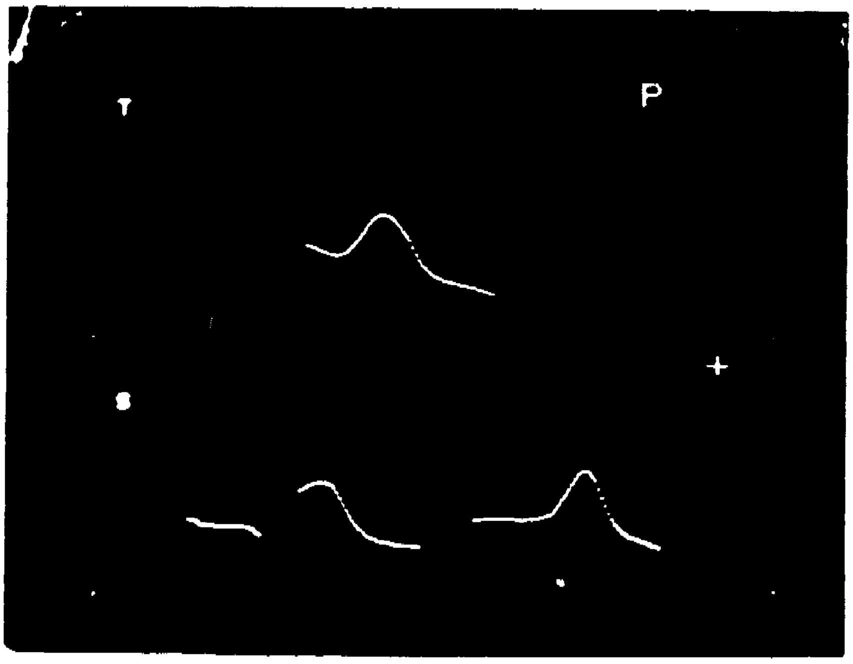


Fig. 7 . Pitch display for two utterances produced by the same male speaker. Top trace: "How are you?" Bottom trace: "I'm fine. How are you?"

which no points are plotted represent either voiceless sounds or silent intervals. This display is used as an aid in the training of control of fundamental frequency, either with sustained sounds or with speech material.

The next two displays shown on the function menu, the high-low and the front-back displays, are illustrative of representations of selected attributes of the spectrum shape as a function of time. The high-low and front-back functions were developed originally for use in a system to aid in the teaching of correct pronunciation to learners of a second language (Kalikow and Swets, 1972). The high-low function is related to the frequency of the first formant, and the front-back function is similarly related to the second-formant frequency. The first- and second-formant frequencies are known to be related to the tongue height and to the front-back position of the tongue hump, respectively (Peterson and Barney, 1952).

Figure 8 shows inferred tongue positions on the high-low dimension as functions of time. The top displays were produced by the vowels in "sod" (left) and "seed" (center) and the diphthong in "side" (right). The corresponding bottom displays were produced by the vowels in "bought" and "boot," and the diphthong in "bout." The front-back displays shown in Figure 9 were produced by "sod," "seed," and "side" (top, left to right), and "pose," "peas," and "poise" (bottom, left to right).

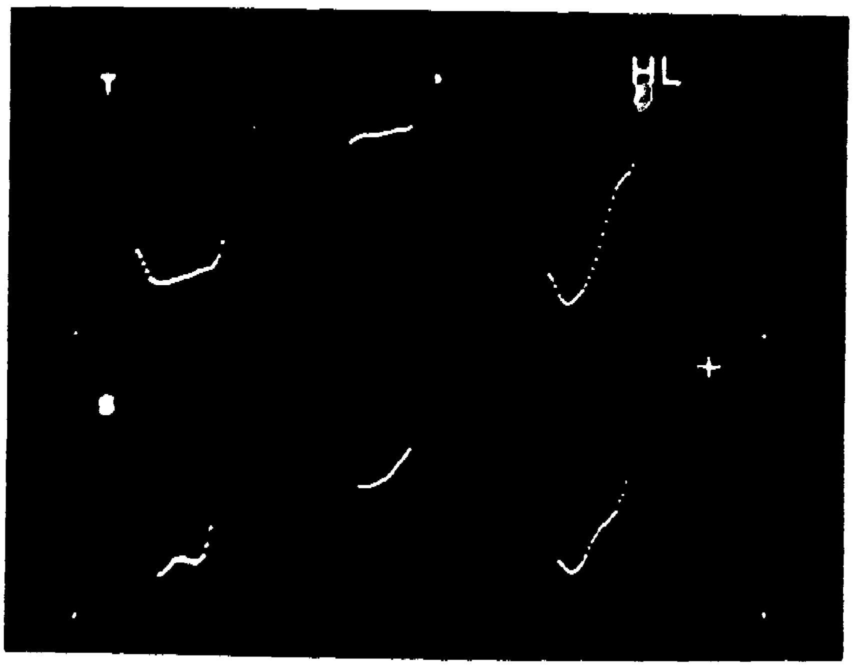


Fig. 8. High-low function display for six words spoken by a male. Left to right; top to bottom: "sod, seed, side"; "bought, boot, bout."

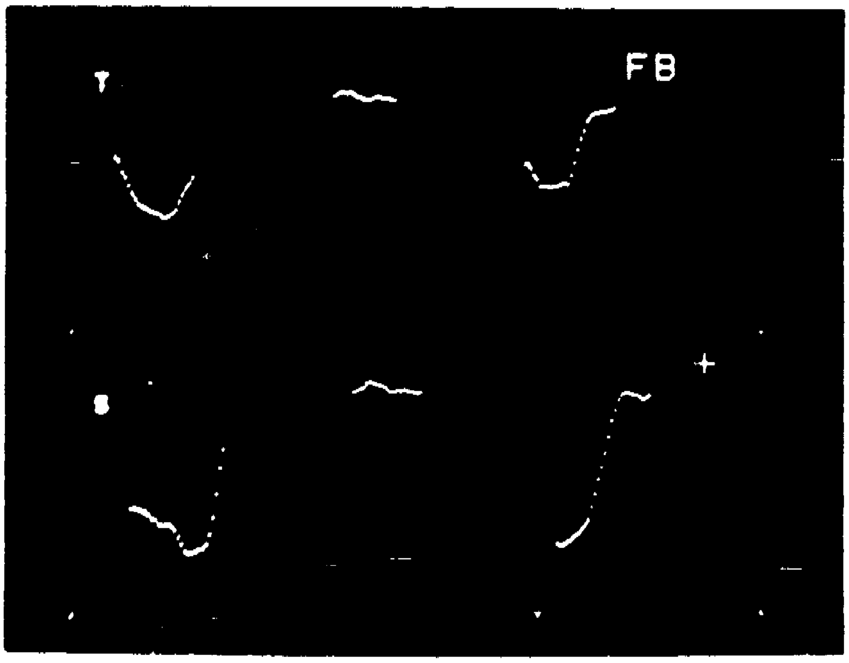


Fig. 9. Front-back vowel function display for six words spoken by a male. Left to right; top to bottom: "Sod, seed, side"; "pose, peas, poise."

While these displays give reasonable approximations to tongue position for nonnasal vowels of normal speakers, they may not give the desired indication of tongue position when the vowel is nasalized, when the spectrum of the glottal output deviated from normal (as in the case of a breathy voice), or when the glottal opening is too large, thus allowing acoustic coupling to the subglottal cavities. Since problems of this type are common in the speech of the deaf, these displays, although they were not developed for this purpose, may be of some use as aids in the training of vowel articulation for deaf students.

The high-low (HL) function has also been found to be useful for indicating "breathy voice quality" for low vowels such as /ɑ/, /ɔ/, /æ/, /ʌ/, all of which have relatively high first-formant frequencies. This function is given by

$$HL = F_1 + F_2 - F_3 - F_4$$

A positive value of HL for a low vowel shows that there is excessive low-frequency energy in the glottal output, and this attribute appears to be correlated with breathy voice quality. A normal, clear voice quality would give a large negative value for HL, for low vowels, since the first-formant spectral peak is in the range of filters 3 and 4. A display of HL is being used on an experimental basis as an aid in training deaf students to improve voice quality.

Another function option is a voicing display, which shows a horizontal line whenever voicing is present, and nothing otherwise.

This display can be used, with suitable utterances, to represent timing and rhythmic properties of speech. Figure 10 gives an example for the words "school teacher." The display shows the temporal properties best with utterances whose syllabic structure is reflected by the interspersing of voiced and unvoiced sounds. Typically, voicing is not shown by itself, but in combination with either loudness or nasality, as will be explained below.

Each of the time-function displays that has been described so far represents a single parameter of speech in isolation. The program provides the user also with several displays of combinations of parameters. Figure 11 is an example of a display that represents pitch and loudness in combination. The upper record represents the sentence "She's a girl"; the bottom one, "She's a tall girl." (The second sentence was spoken without special emphasis on "tall.") Two traces are associated with each voiced segment of an utterance; the bottom trace represents pitch and the top one loudness. The loudness function is plotted relative to the pitch trace; that is to say, each point on the loudness function is the sum of the pitch and loudness functions at that point in time. Unvoiced sounds are represented by only a single trace, which is the loudness function. Thus, in the lower record in Figure 11, one can distinguish two voiced and two unvoiced segments. The two unvoiced segments correspond to the fricative consonant /s/ with which the utterance begins and the voiceless /t/ at the beginning of the word "tall."

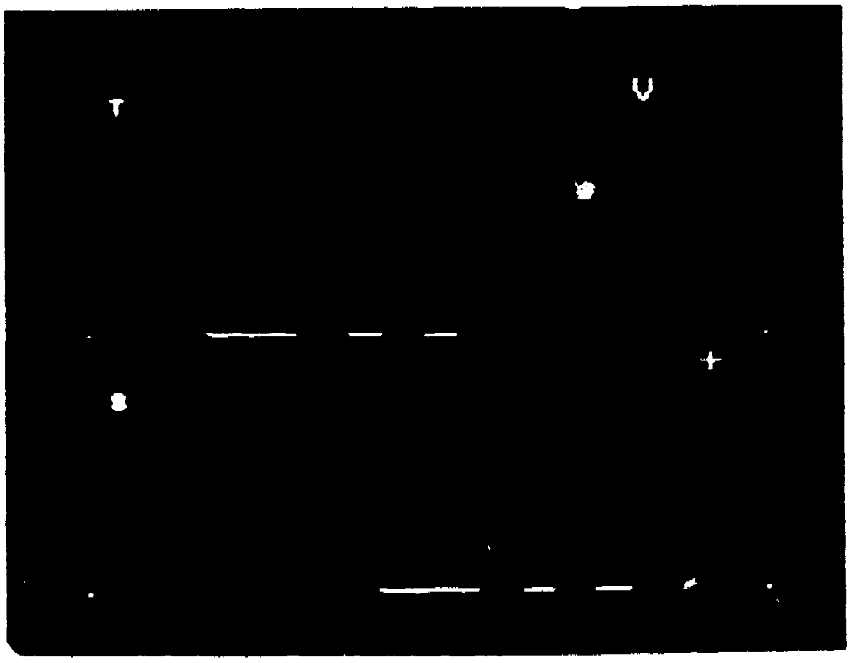


Fig. 10 Voicing function display for a male speaker's two utterances of the 3-syllable phrase "school teacher."

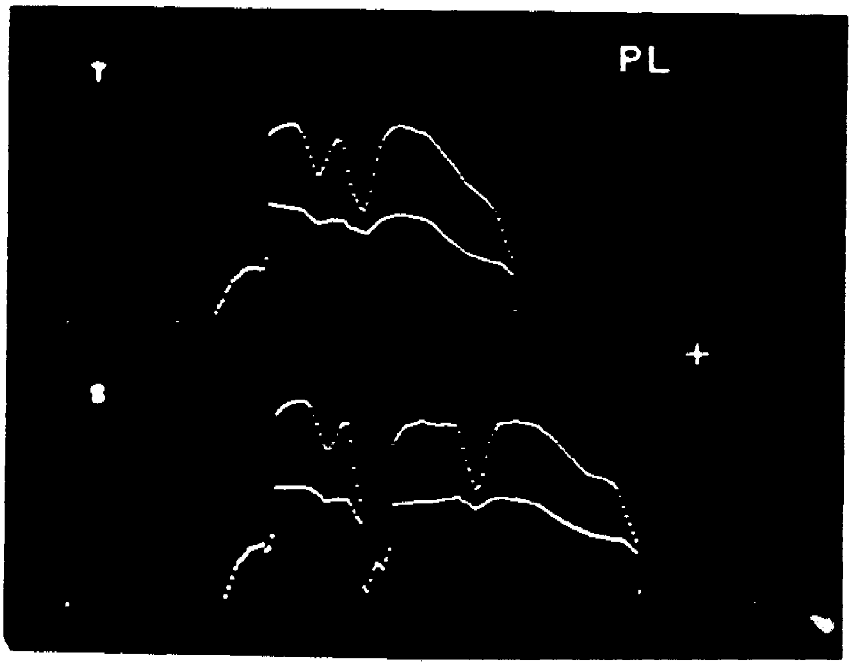


Fig. 11. Pitch-loudness composite display for a male speaker's utterance of two sentences. Top: "She's a girl." Bottom: "She's a tall girl." See text.

The display shown in Figure 12 is similar to that shown in Figure 11 except that here voicing instead of pitch is represented in combination with loudness. In this case, the words that are represented are "destitute" (upper) and "testify" (lower). Both of these displays (Figures 11 and 12) can be used to show patterns of timing and stress. Experience to date suggests that timing information is more readily conveyed by the voicing-loudness composite (Figure 12). This display is somewhat less "busy" than that shown in Figure 11, and the temporal structure of the utterance may be equally apparent. The pitch-loudness composite conveys some information about stress and intonation, however, that is not contained in the voicing-loudness composite. The acoustic correlates of stress are not well understood, but they are known to include variations in timing, loudness, and pitch. The displays in Figure 11 contain some information on each of these factors: the relative loudness of the stressed syllables, the changes in pitch from one syllable to the next and within each syllable, and the relative brevity of the unstressed medial vowel are all visible in these displays. Figure 6 also demonstrates the interaction of these variables in determining two different stress patterns on the same sentence. From top to bottom the traces represent loudness, pitch, and pitch-loudness composite, respectively.

The pitch-loudness and voicing-loudness composite displays can also be used to assist the training of articulation. Missing or

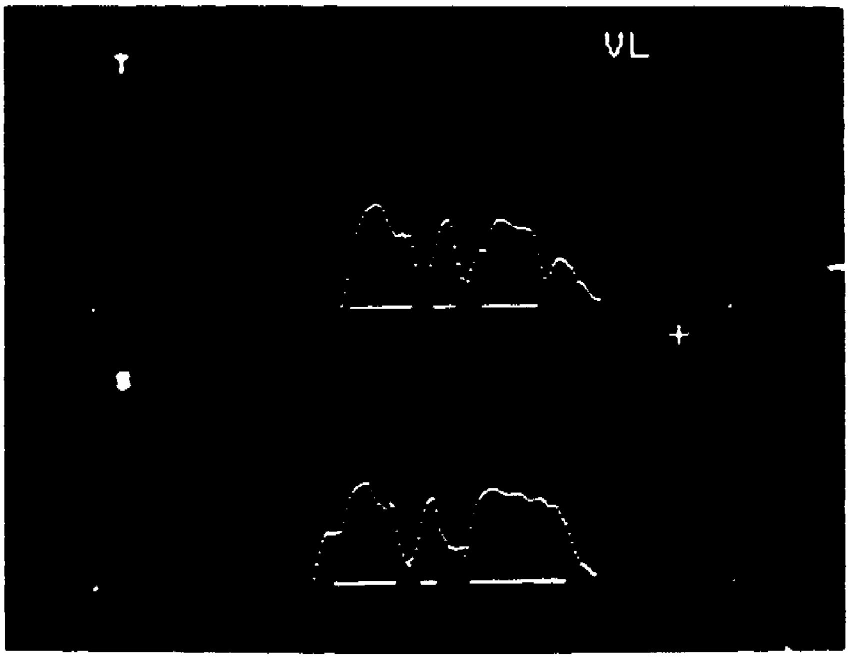


Fig. 12. Voicing-loudness composite display for a male speaker's utterances of two 3-syllable words. Top: "destitute"; bottom: "testify." Note breaks in the loudness function caused by total vocal-tract closure for the three medial occurrences of the /t/ in the two words.

adventitious sounds, and in some cases speech-sound substitutions, can be seen fairly clearly in these displays. Confusions between voiced and voiceless speech sounds are particularly easy to detect.

The displays shown in Figure 13 combine representations of voicing and of nasality. The method that is used to obtain nasality information has been described by Stevens, Kalikow, and Willemain (1974). The function that represents the degree of nasality is simply the rectified, smoothed, and log-converted output of the accelerometer attached to the nose. Some data on this parameter from normal and hearing-impaired speakers producing utterances containing nasal and nonnasal sounds have been collected (Stevens, Nickerson, Boothroyd, and Rollins, in preparation), and have been organized in a form that can be used to set criteria for nasality in a training situation. Suffice it to say here that, whereas velar control has long been recognized to be a problem for many deaf speakers, not many attempts have been made to obtain objective measures of nasality that could be used in real-time speech-training displays. The nasality function illustrated in Figure 13 does distinguish between nasal and nonnasal sounds produced by speakers with normal hearing. The traces shown in the figure represent the words "Monday" and "Tuesday." The nasal peaks associated with the m and n are clearly identified, and of course the intervening vowel is also highly nasalized compared to the other vowels in the sample. (The reason for the horizontal dotted lines in this figure is discussed below.)

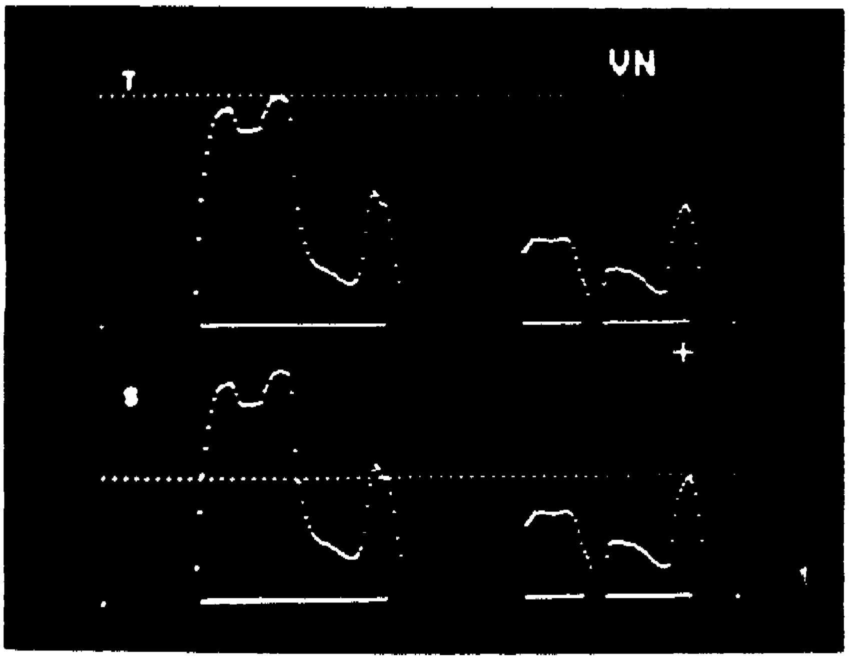


Fig. 13. Voicing-nasality composite display for a normal male speaker's utterance of "Monday" and "Tuesday." Data from the lower half of the display have been copied to the upper half under user control.

For some purposes one may wish to think of nasality as a binary attribute. The displays shown in Figure 14 illustrate the possibility of doing this with the time-plot program. These displays represent voicing and nasality in combination, but both are represented as binary attributes. A single horizontal line indicates voicing; double horizontal lines bracketing the voicing line indicate nasality. The latter lines appear whenever the nasality function exceeds a criterion that can be set by the user of the system. The utterance that is represented in both areas of Figure 14 is the sequence of words "one, two, three." The upper display was obtained by setting the nasality criterion relatively high (278 on an arbitrary scale from 0 to 400). In this case, nasality is indicated only for the single nasal consonant that occurs in the utterance, the *n* in "one." The lower display was produced by decreasing the nasality criterion to 156 (about 12 dB lower than the criterion used in the upper display). In this case, the entire "one" is considered nasalized, and brief nasal segments are identified at the ends of the words "two" and "three." (A rise in nasality at the ends of isolated words ending in vowels has been observed in many of our samples of speech of normal-hearing speakers. For example, Figure 15 shows the nasality functions for the utterance represented in Figure 14. This sample was produced by a male speaker with normal hearing.)

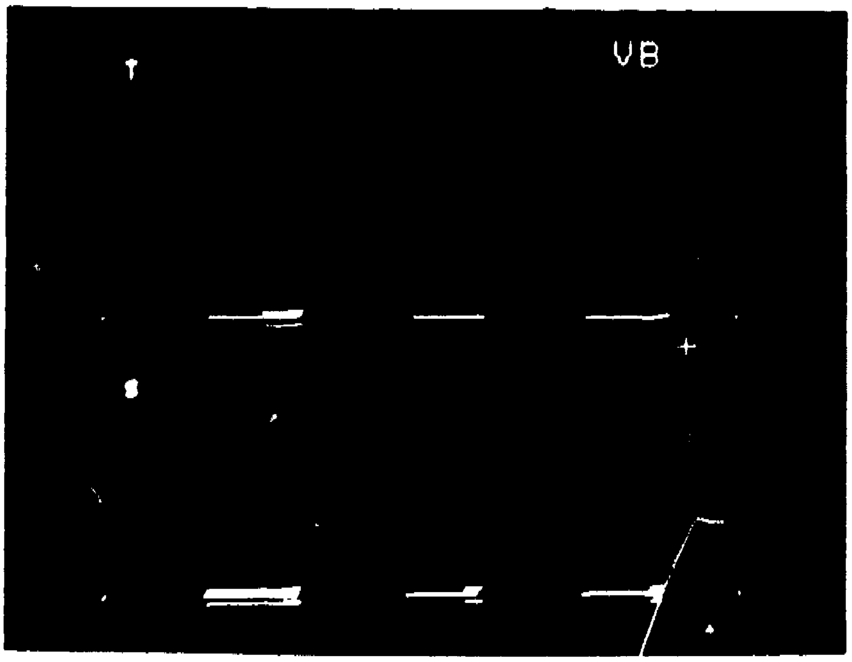


Fig. 14. Voicing-binary nasality composite display for a normal male speaker's two versions of the utterance "one two three." Differing nasality criteria were employed in the two halves of the display. See text.

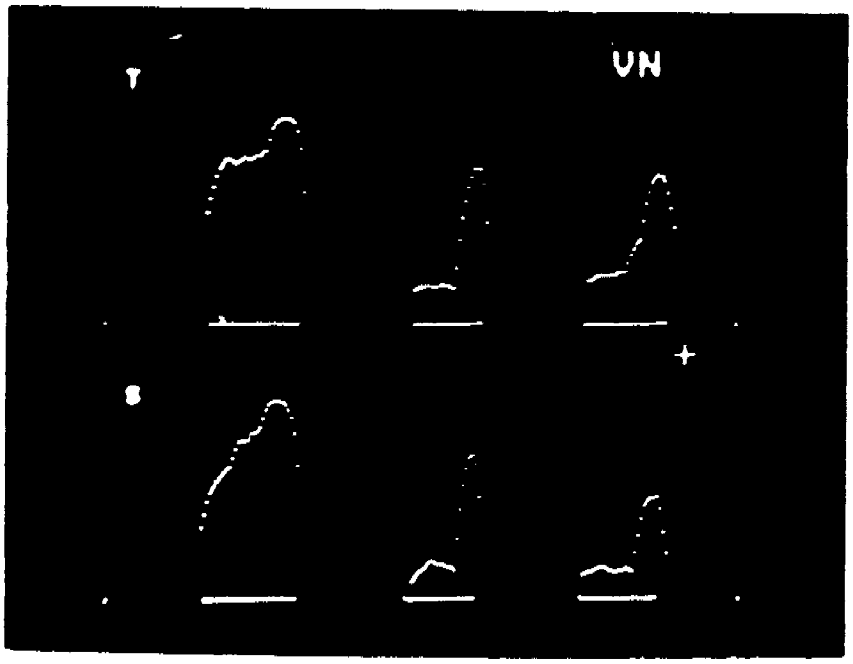


Fig. 15. Voicing-nasality display for a normal male speaker's two versions of the utterance "one two three." These data served as the basis for the display photographed in Fig. 14 , after the selection of nasality criteria as described in the text.

Another option provided by the function menu in the time-plot program is for a composite display of three features: voicing, loudness, and a binary indication of nasality. The display is illustrated for the utterance "one, two, three" in Figure 16. The nasality criteria for the upper and lower displays were the same as in Figure 14.

There is little doubt that the features represented in the various function displays that have been described play important roles in the production and understanding of speech. There is a degree of arbitrariness, however, concerning the ways in which these features have been encoded in those displays. The use of time functions has a certain face validity inasmuch as speech is describable in terms of a set of time-varying parameters. Moreover, representing time spatially means that the display has a "memory," which makes it relatively easy to distinguish visually some short-lived aspects of an utterance that might be difficult to detect in a display that represented properties of the speech only for the extent of their duration in real time. For example, the fleeting initial plosive in "testify" would be difficult to distinguish from the initial voiced consonant in "destitute" in a continuously changing real-time display with no memory, such as the spectral representation illustrated in Figure 3. The difference is easy to see, however, in the time plots shown in Figure 12. A

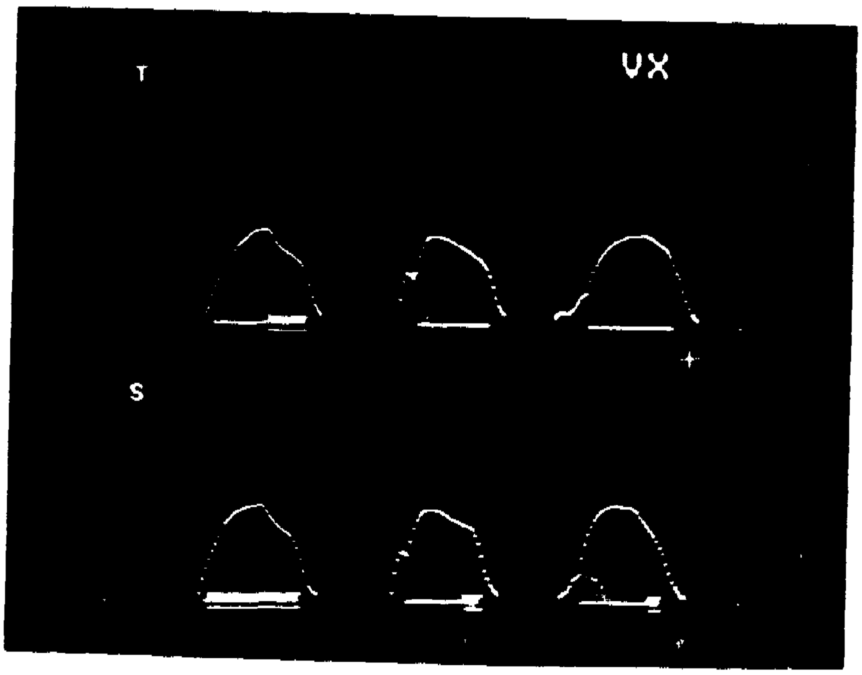


Fig. 16. Voicing-loudness-binary nasality composite display for a normal male speaker's two versions of the utterance "pne two three." Nasality criteria identical to those used in Figure 14.

disadvantage in the use of time functions is the fact that many features cannot easily be represented simultaneously in an integrated fashion. Showing several time functions in parallel on the same display is a possibility; it is not clear, however, that the viewer can make effective use of such a display. Considerable experimentation will undoubtedly be required not only to determine the relative utility of various types of information for speech training, but also to evolve optimal ways of encoding that information for visual (or other types of) presentation.

Display mode options. Each of the function displays discussed in the preceding paragraphs can be operated in each of three modes. The user selects the mode which he wishes to use by means of a menu display in the same way that he selects a function. In graph mode, the display is generated from left to right as the speaker speaks. The "now" position moves from left to right at a rate of about 9.3 centimeters per second. When the display drops off the right-hand side of the screen, it "wraps around" and reappears on the left-hand side. This is the conventional mode of operation of standard oscilloscope displays. One disadvantage of this mode of operation is that if one wants to monitor the instantaneous changes in the function as the speaker speaks, one must move one's eyes from left to right to follow the developing trace, and periodically move quickly back to the left to pick it up when it wraps around. A further disadvantage is the awkwardness of the mode for utilizing the freeze option. Suppose, for example, that the display is frozen when the "now" point is about midway across the scope. The most recent part of the sample will be to the left of that point, whereas the oldest part of the sample will be to the right of it; and this seems like an undesirable arrangement. To avoid this problem, the displayed function is justified to the right edge of the scope whenever the freeze option is invoked, but this means that invoking the freeze option causes an abrupt change in the display format that could be confusing to a child. (An alternative graph-mode procedure that is sometimes used, though not in this system, is to erase the entire display whenever wrap-around occurs. This procedure makes it difficult to capture speech samples on an ad hoc basis, however, inasmuch as the beginning of the utterance must be synchronized with the beginning of the sweep of the display from the left-hand margin.)

We refer to the second mode of operation as <u>flow</u> mode. In this case, the "now" point is fixed at the right-hand side of the display and the past moves off to the left as the speech is produced. The graph-mode display operates as though one were generating the functions by moving a pencil from left to right; the flow-mode display operates as though one were simply moving the pencil up and down while moving the paper from right to left. Viewing the latter display is similar to watching a strip-chart recorder through a window of fixed size. In this case one need not move one's eyes laterally to attend to the instantaneous changes in the display as the speech is produced, nor to refixate from one side of the display to the other to accommodate wrap-around. A disadvantage is that the programming demands are somewhat greater inasmuch as each updating of the display requires a repositioning of all the points as opposed to the addition of a single point.

Our experience with the system to date indicates that the flow-mode presentation is strongly preferred to graph-mode by users of the system. Anderson (1960) was perhaps the first to make use of this type of representation in a speech-training aid. He designed a pitch indicator in which pitch was displayed on a rotating cathode ray tube. The tube was masked except for a 60° arc at the top center, and the electron beam was moved up and down at the right edge of the window in the mask as the tube rotated counterclockwise. Thus, a trace of frequency versus time was generated within the window. In describing his indicator, Anderson remarked on the desirability of being able to "freeze" the display on demand.

The third option provided by the display-mode menu is for a _delay_ mode of presentation. In this case the function is not displayed while it is being generated (as is the case for both graph and flow modes) but only on demand. When the delay mode has been invoked, the display operates in the following way. Whenever the user wants to capture an utterance, he presses a designated button. The computer at that point stops sampling speech and retains the record for the immediately preceding two seconds. By pressing another designated button, the user can then request that the captured sample be displayed whenever he wishes.

The purpose for the delay-mode option is to make it possible for the student to attempt to produce an utterance without the benefit of watching the developing display, or to evaluate his own effort to accomplish some training goal before seeing a representation of his speech. The reason for using this approach is to attempt to teach the child to rely on proprioceptive cues for judging his performance and thereby hopefully to facilitate transfer of what he learns in the speech laboratory to his everyday speech. After having made his assessment, the display can be shown, and used to confirm or infirm his own evaluation.

Additional features of the time-plot program. A feature that has recently been added to the time-plot program enables the user to set criterion levels or to measure the level of any given function at particular points. This feature works in the following way. Two control knobs determine the heights of two horizontal lines across the face of the display. One of these lines is associated with the teacher's display region (upper half of scope), and the other with the student's region (lower half of scope); both work in the same way. Each line extends all the way across the display and its height is moved up and down by turning the associated control knob clockwise and counterclockwise, respectively. The line can be made to disappear by turning the knob clockwise as far as it will go. The height of the lines may be determined at any time by pressing the "knobs" button. For example, if this button is pressed when the lines are set as shown in Figure 17a, the display shown in Figure 17b would appear. The numbers here represent the heights of the lines in terms of an arbitrary scale that goes from 0 to 400. The only exception to this rule is the pitch display, in which case the heights of the lines are reported in Hz. The utterance represented in Figure 17 is "Bob got the shoe." The upper trace in each frame was generated by emphasizing "shoe," the lower one by emphasizing "Bob." The parameter that is displayed is pitch.

To use the lines as criteria, the teacher may adjust them to a position that represents the level of some parameter to which the student is to relate his speech. For example, if the student's problem is hypernasality, the teacher might adjust a line for use with the nasality display and ask the student to attempt to keep the nasality trace below that line while producing speech sounds that should not be nasalized. Or, if the problem is falsetto voice, a similar procedure might be followed with the pitch display. The teacher may, of course, adjust the criterion to more and more demanding levels as the student acquires more skill at the task.

Use of the criterion lines in the voicing-nasality display has been illustrated in Figure 13. In this example, to determine the difference between the nasality peaks obtained in the utterances "Monday" and "Tuesday," one would adjust the line first to the

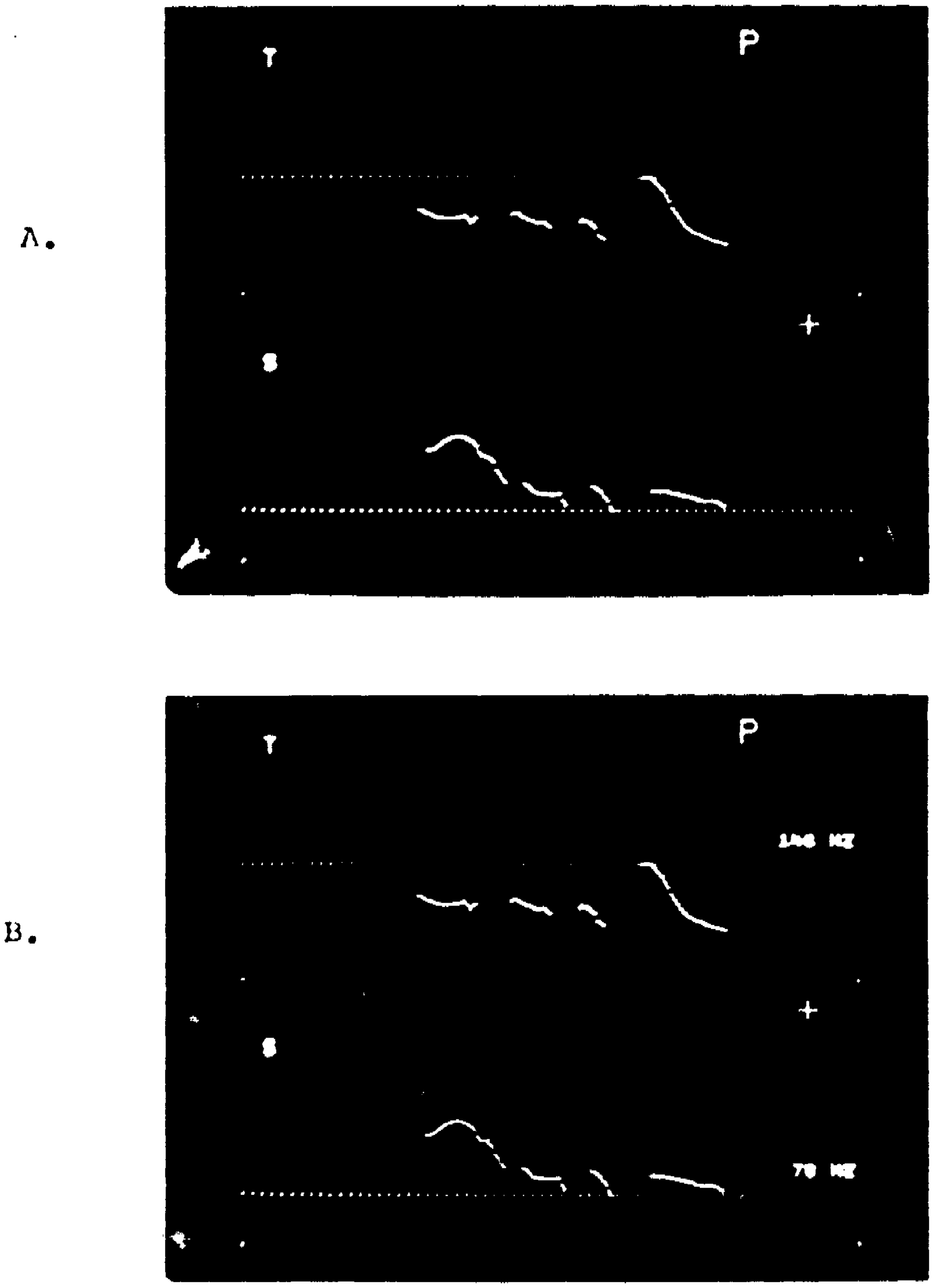


Fig. 17. Pitch display illustrating criterion level readout capability. See text.

part of the nasality function representing the M in "Monday," and then to the highest point in the function representing "Tuesday," and obtain a reading for each. In the case of this example, the readings were, respectively, 375 and 199.

In addition to the horizontal lines the user can also display and adjust two vertical lines, either to set criteria for timing or to take duration measurements from a displayed trace. These lines were provided particularly to facilitate training with respect to timing. The sequence of photos in Figure 18 illustrates how these lines may be used. In the first frame the lines have been adjusted to bracket a loudness representation of a teacher's utterance "at the park." The number (.82), which appears in response to a press of a button, indicates the duration of the bracketed time in seconds. The second frame represents a student's attempt to make the same utterance. (Both "teacher" and "student" were normal-hearing speakers, and the student intentionally spoke relatively slowly for the sake of the illustration.) In the third frame, the student's trace has been shifted to the left so that its beginning is lined up with that of the trace produced by the teacher, and the right vertical line has been moved so that the two lines now bracket the student's trace; the number (1.05) indicates the duration of his utterance. The student's trace can be shifted right or left any desired amount under push-button control.

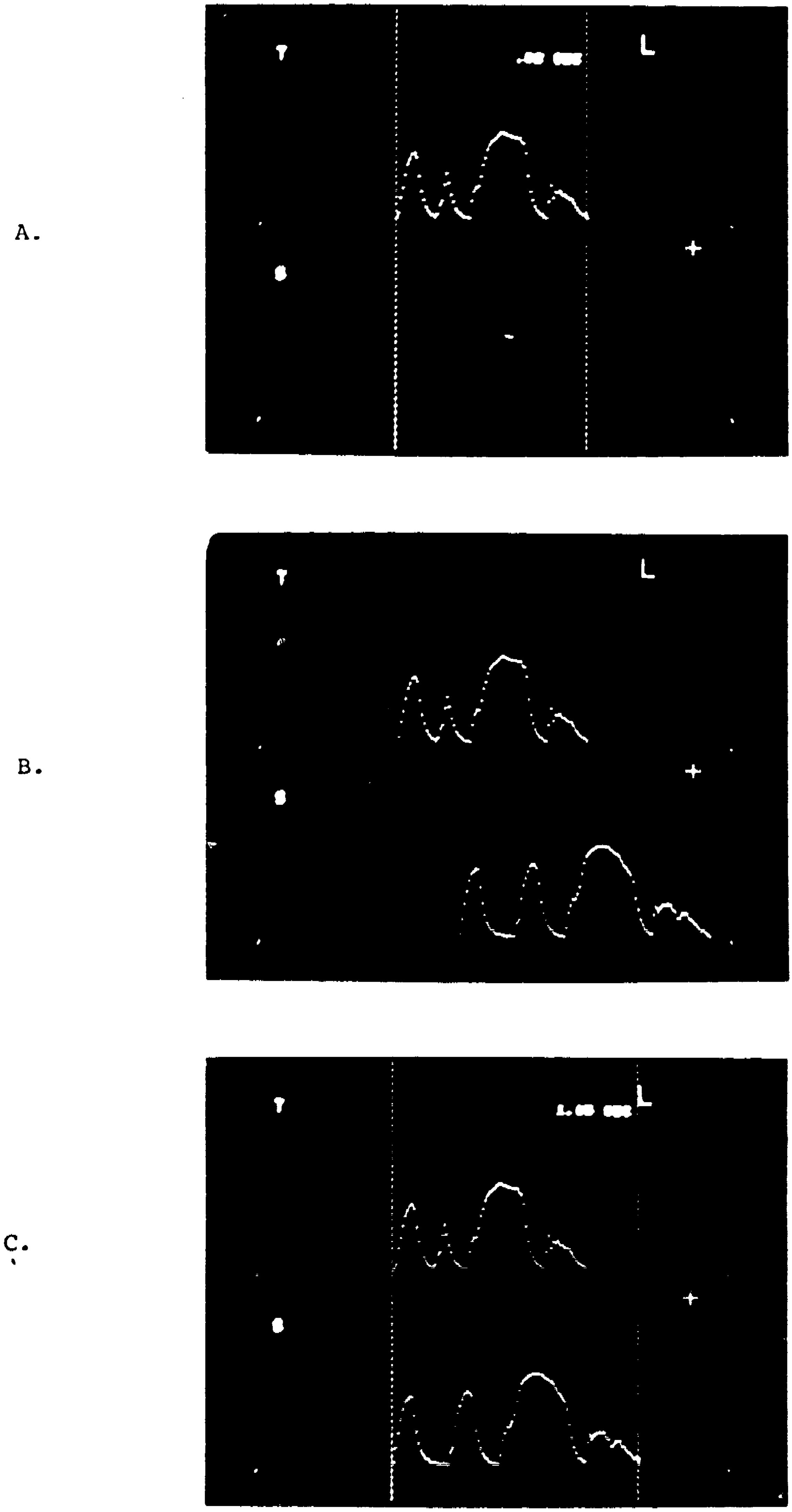


Fig. 18. Loudness displays of the utterance "at the park." A: Teacher's trace bracketed by adjustable lines .82 sec. apart. B: Student's attempt to match teacher's trace. (Both teacher and student were normal-hearing males and the latter intentionally spoke slowly.) C: The student's trace is moved to the left so its beginning coincides with that of the teacher's trace, and the lines are adjusted to bracket the student's trace. (In order to photograph the dimmer vertical lines, the intensity of the entire display had to be increased to the point of blurring the numbers.)

The features described in this section were added to the time-plot program as a result of suggestions made by the teachers who are using the system. More generally, the entire evolution of the system has been guided by a constant exchange of ideas between developers and users.

THE POSSIBILITY OF AUTOMATED EVALUATION CAPABILITIES

Visual displays of the sort that have been described in this report provide the speaker with feedback concerning the results of his efforts to make certain speech sounds. If the speaker is to use that feedback in order to improve his efforts, he must be able to interpret it. In particular, he must know whether the display in any given instance is a satisfactory one. Given the capabilities of the system as described above, this knowledge is gained in one of two ways: either the tutor passes judgment on a display and tells the student whether it is satisfactory or not, or the student compares the display that he has produced with a target pattern and makes the judgment visually himself. It is not always easy to judge, however, whether a display _is_ acceptable or not. Not only must the teacher, or the student, learn to ignore irrelevant aspects of the patterns that are produced, but he must have a good understanding of the ways in which fluent speech varies with respect to the features that are represented by the different displays. It would be most helpful if it were possible to provide the computer itself with the capability of evaluating at least some of the displays objectively.

An attempt was made to implement such a capability for the purpose of assessing the acceptability of some of the gross temporal aspects of an utterance. The capability was programmed for displays of voicing (voicing, voicing-loudness, voicing-nasality, voicing-binary nasality, and voicing-loudness-binary nasality composites). The procedure compares the utterance represented in the bottom half of the display against the one that is represented in the top half of the display with respect to the voiced components of that utterance. Specifically, it performs a series of tests as follows: (a) It compares the two utterances with respect to their overall duration. By overall duration, we mean the time from the beginning of the first voiced segment to the end of the last voiced segment. If the duration of the student's utterance equals the duration of the standard, plus or minus x%, test b is performed. If it does not equal the standard within the specified tolerances, then the words "too long," or "too short," as appropriate, appear under the display of the student's utterance. (b) The voicing segments are counted, and if the number of voiced segments in the student's utterance equals the number of voiced segments in the standard utterance, test c is performed. If the numbers are not equal, then the words "too many breaks," or "too few breaks," as appropriate, appear on the display of the student's utterance. (c) The durations of the individual voicing segments in the student's utterance are compared to the corresponding segments in the standard utterance, and if, in each case, the duration of the segment of the student's utterance equals that of the standard, plus or minus x%, then the word "good" appears under the student's display. If the duration of one or more of the segments is not within the specified tolerances, the words "incorrect timing" appear under the student's display. The value of the parameter x, which is used in tests a and c, is adjustable by the teacher. Presumably, one would begin with a relatively large value of this parameter and then decrease it regularly, as the student becomes more proficient at the task.

This evaluation feature has not yet been used enough to justify an opinion concerning whether it will be helpful. It may well be that we will discover that, at least in the case of timing, such an approach is not feasible. At best, the procedure will undoubtedly have to be refined. The general desirability of objective evaluation procedures, however, seems clear, and while their development will certainly involve considerable trial-and-error exploration, their potential benefit should justify the effort. One of the most apparent advantages of having such procedures is the fact that their existence would greatly enhance the use of displays by students themselves in self-tutoring sessions.

CONCLUDING COMMENT

We stress again that the displays that are described in this report are still in the process of being developed. We have only begun to collect evaluative data, and so are not in a position to assert their effectiveness. Our initial experience with the system has been encouraging; however, it seems clear that how effective any speech-training aids will prove to be in practice will be bounded above by the specifics of the ways in which they are used. As technical developments make it feasible to do increasingly complex real-time analyses of speech and to generate nearly anything one wants by way of displays, it becomes more and more apparent that pedagogical uncertainties impose the real limits on what one can expect to accomplish with speech-training aids, no matter how technologically sophisticated they may be.

A thought experiment demonstrates this point. Imagine a machine that could perform in real time any type of analysis of speech that one wished, and generate any display that one might specify. The fact is that w do not really know what analyses should be performed or what displays should be developed. Moreover, even if we knew the answer to these questions, it is not clear that enough is known about speech acquisition among the deaf to provide the basis for the training procedures that would take full advantage of such capabilities. What does seem clear to us is that the flexibility of a computer-based system provides opportunities for the type of exploration that is likely to be required to make progress on these problems.

Finally, the sort of close collaboration between researchers and teachers that we have attempted to maintain in this project is essential, we believe, if efforts to evolve effective training aids are to have a reasonable chance of success. This is not a new idea. Kopp (1938) expressed the need for a greater interaction between teachers and researchers by suggesting that the field would benefit "if we could make more teachers researchers, and more researchers teachers." Other writers have also advocated such interaction (Borrild, 1968; Denes, 1968), but few serious attempts to collaborate seem to have been made. The strategy is a reasonable one, we feel, not only for the development of this particular system but for that of any complex system that is to involve a real-time interaction between men and computers on problems for which approaches are not highly formalized and the solutions are not well understood. As David (1962) has pointed out, the great versatility of the computer represents both an opportunity and a challenge. The opportunity is for creativity and innovation; the challenge is to be discriminating and practical. A close coupling between a system's developers and its users is perhaps the only way to assure a balance between innovativeness and practicality from which something both new and useful may emerge.

NOTES

1. All of the figures representing displays in this report were made by taking Polaroid snapshots of a "slave" oscilloscope that shows a duplicate, in miniature, of the display at the teacher-student station.

ACKNOWLEDGMENT

Each of the following individuals has contributed significantly to the design, implementation, and use of the system described in this report: Robb Adams, Patricia Archambault, Arthur Boothroyd, Douglas Dodds, Ann Rollins, Robert.Storm, and Thomas Willemain.

This project was sponsored by the U.S. Office of Education Media Services and Captioned Filsm Branch of the Bureau of Education for the Handicapped, under Contract No. OEC-0-71-4670(615). It is a pleasure to acknowledge the encouragement and helpful suggestions of Lois Elliott, who served as contract monitor during the initial stages of the project.

Anderson, F. An experimental pitch indicator for training deaf scholars. *Journal of the Acoustical Society of America*, 1960, *32*, 1065-1074.

Borrild, K. Experience with the design and use of technical aids for the training of deaf and hard of hearing children. *American Annals of the Deaf*, 1968, *113*, 168-177.

David, E. E., Jr. Speech in the computer age. *Volta Review*, 1962, *64*, 394-397.

Denes, P. D. Speech science and the deaf. *Volta Review*, 1968, *70*, 603-607.

Kalikow, D. N. and Swets, J. A. Experiments with computer-controlled displays in second-language learning. *IEEE Transactions on Audio and Electroacoustics*, 1972, *AU-20* 23-28.

Kopp, G. A. The application of recent findings in the field of speech correction. *Volta Review*, 1938, *40*, 638-640.

Levitt, H. Speech processing aids for the deaf. *IEEE Transactions on Audio and Electroacoustics*, 1973, *AU-21*, 269-273.

Nickerson, R. S. and Stevens, K. N. An experimental computer-based system of speech training aids for the deaf. In *Proceedings, Conference on Speech Communication and Processing*, Newton, Mass., April 1972, pp. 238-241.

Nickerson, R. S. and Stevens, K. N. Teaching speech to the deaf: Can a computer help? *IEEE Transactions on Audio and Electroacoustics*, 1973, *AU-21*, 445-455.

Peterson, G. E. and Barney, H. L. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 1952, *24*, 175-184.

Pickett, J. M. Recent research on speech-analyzing aids for the deaf. *IEEE Transactions on Audio and Electroacoustics*, 1968, *AU-16*, 227-234.

Stevens, K. N., Kalikow, D. N., and Willemain, T. R. The use of a miniature accelerometer for detecting glottal waveforms and nasality. BBN Report No. 2907, Sept. 1974 to U.S. Office of Education under Contract No. OEC-0-71-4670(615).

Stevens, K. N., Nickerson, R. S., Boothroyd, A., and Rollins, A. Assessment of nasality in the speech of deaf children. BBN Report No. 2902, Sept. 1974 to U.S. Office of Education under Contract No. OEC-0-71-4670(615).