E R R A T A T O

UIUCDCS-R-71-479

COO-2118-0024

A COMPARATIVE STUDY OF SOME VISUAL
SPEECH DISPLAYS

BY

Bernard J. Nordmann Jr.

September 10, 1971

Title Page

Footnote should read:

This work was supported in part by the Atomic Energy Commission under Contract AT(11-1)-1018 and AT(11-1)-2118 held by the Digital Computer Laboratory, University of Illinois, and in part by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract DAAB-07-67-C-0199, held by the Coordinated Science Laboratory, University of Illinois.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

UIUCDCS-R-71-479

A COMPARATIVE STUDY OF SOME VISUAL
SPEECH DISPLAYS *

BY

Bernard J. Nordmann Jr.

September 10, 1971

Department of Computer Science
University of Illinois
Urbana, Illinois 61801

* This work was supported by Contract AT(11-1)-1018 with the U.S. Atomic Energy Commission through September 30, 1970. Current support is under Contract AT(11-1)-2118 with the above agency.

A COMPARATIVE STUDY OF SOME VISUAL
SPEECH DISPLAYS

BY

BERNARD JOSEPH NORDMANN JR.

S.B., Massachusetts Institute of Technology, 1966

S.M., Massachusetts Institute of Technology, 1966

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1971

Urbana, Illinois

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

ACKNOWLEDGMENT

The author would like to express his sincere gratitude to Professor Sylvian R. Ray and Professor H. Gene Slottow for their advice and encouragement with this thesis.

Thanks are also due to Mrs. Judy Arter and Mrs. Roberta Andre' for typing and correcting the manuscript and for their continuous requests for manuscript to type which was a major factor in the completion of this thesis before the deadline. Last but not least, thanks are due to Mr. Stan Zundo and Tom Gedgaudas for preparing the figures and tables used in this thesis.

A COMPARATIVE STUDY OF SOME VISUAL SPEECH DISPLAYS

Bernard Joseph Nordmann Jr., Ph.D.
Department of Computer Science
University of Illinois at Urbana-Champaign, 1971

The purpose of the present project was to develop a computer speech display simulation system capable of generating a wide variety of speech displays from a recorded speech input. Eventually it is hoped that this will lead to a system whereby a person can obtain visual feedback as a corrective measure for word pronunciation. The basic system would involve two displays, one representing the subject's pronunciation of a particular word and the other representing a correct pronunciation of the word. A computer would be used to process the incoming speech and produce a display containing features highly relevant to correct pronunciation. The subject's task would be to detect differences in the two displays and to change his pronunciation so as to make them more similar.

After conducting an extensive literature search to determine the types of schemes which had previously been used to display speech sounds, a basic interactive display system was programmed using the CSL's CDC 1604 computer-graphics facility. The system has been designed to be open-ended and currently can produce photographs of a variety of display types. Unfortunately, the system as it stands now cannot operate in real time due to the slowness of the CDC 1604.

The simulation system was used to produce examples of several different types of displays. These displays were used in a series of preliminary tests designed to develop techniques for comparing the effectiveness of various types of displays. Several corrections and refinements to the testing methods are discussed.

TABLE OF CONTENTS

Chapter	Page
1. INTRODUCTION	1
2. CHARACTERISTICS OF SPEECH	5
2.1 Problems in Speech Analysis	5
2.2 Significant Parameters of Speech	7
3. HISTORY OF SPEECH DISPLAYS	12
3.1 Early Displays	12
3.2 Spectrographic Displays	13
3.3 Spectrographic Variations	15
3.4 Other Linear Time Displays	16
3.5 Two-Dimensional X-Y Displays	17
3.6 Zero-Crossing Displays	20
3.7 Pitch Extracting Displays	22
3.8 Miscellaneous Formats	23
3.9 The Use of Speech Displays	24
4. PROPOSED STUDY	29
4.1 Outline of the Study	30
4.2 Theoretical Significance of the Comparison Tests	32
5. DISPLAY DESCRIPTIONS	36
5.1 Variable-Intensity TV Scan Display	36
5.2 Continuous Line Display	37
5.3 Spectrogram	37
5.4 Formant Extracting Display	42
5.5 Zero-Crossing Display	46
5.6 Zero-Crossing vs Amplitude Envelope	46

	Page
6. SPEECH DISPLAY SIMULATION SYSTEM	52
6.1 The Common Data Base	52
6.2 The Command Processor	55
6.3 The Speech Display Routines	57
6.4 The Subprocessing Routines	59
6.5 Basic System Principles	60
7. RESULTS	62
7.1 Recordings	63
7.2 Data from the First Test	65
7.3 Data from the Second Test	85
8. SUMMARY AND CONCLUSIONS	93
8.1 Comments on the Tests Which Were Performed	93
8.2 Comments on the General Method	94
8.3 Summary	95
REFERENCES	97
VITA	106

LIST OF FIGURES

Figures	Page
1 Effect of Variations in High Frequency Emphasis and Intensity Truncation Using the Word "Shod"	39
2 Effect of Variations in Time Slice Size	40
3 Examples of the Spectrographic Display with Nominal Parameter Values	41
4 Effect of the Peak-Picking Process on the Spectrum Analysis of a Single Time Slice	43
5 Effect of the Peak-Picking Process on the Full Spectrographic Analysis of the Word "Beat"	44
6 Examples of the Formant Extracting Display	45
7 Examples of the Zero-Crossing Display	47
8 Block Diagram for Z_1 and Z_2 vs. Amplitude Envelope Display	48
9 Examples of the Z_1 vs. Amplitude Envelope Display	50
10 Examples of the Z_2 vs. Amplitude Envelope Display	51
11 Relationship Between ISAMP and ISAMPB	53

LIST OF TABLES

Tables	Page
1 Distinctive Features	10
2 Commands Executed by Speech System	58
3 List of Recorded Words	64
4 Learning Rates for Spectrographic Display	67
5 Learning Rates for Zero-crossing Display	68
6 Learning Rates for Formant Extracting Display	69
7 Confusion Matrix for Subject A, Test 1a, Spectrographic Display	70
8 Confusion Matrix for Subject B, Test 1a, Spectrographic Display	71
9 Confusion Matrix for Subject C, Test 1a, Spectrographic Display	72
10 Confusion Matrix for Subject D, Test 1a, Spectrographic Display	73
11 Confusion Matrix for Subject E, Test 1a, Spectrographic Display	74
12 Confusion Matrix for Subject A, Test 1b, Spectrographic Display	75
13 Confusion Matrix for Subject B, Test 1b, Spectrographic Display	76
14 Confusion Matrix for Subject A, Test 1a, Zero-crossing Display	77
15 Confusion Matrix for Subject B, Test 1a, Zero-crossing Display	78
16 Confusion Matrix for Subject A, Test 1b, Zero-crossing Display	79
17 Confusion Matrix for Subject B, Test 1b, Zero-crossing Display	80
18 Confusion Matrix for Subject A, Test 1a, Formant Extraction	81
19 Confusion Matrix for Subject A, Test 1b, Formant Extraction	82

20	Detailed Comparison Matrix for Subject A, Test 2, Spectrographic Display	87
21	Summary Comparison Matrix for Subject A, Test 2, Spectrographic Display	88
22	Detailed Comparison Matrix for Subject A, Test 2, Zero-crossing Display	89
23	Summary Comparison Matrix for Subject A, Test 2, Zero-crossing Display	90

Chapter 1

INTRODUCTION

The purpose of this study is to investigate several methods for producing visual displays of speech signals. Visual speech displays are generally used either as speech analyzers or as speech recognizers. In the first case they can be used to extract a greater or lesser amount of information from a speech utterance and this information can then be recorded and compared with displays of other utterances to determine the types of information which characterize speech. Traditionally, there have been two separate approaches to speech display analysis: one which attempts to determine a display transform which will present all the information necessary to determine the various phonemes and the other which takes a display of a single type of speech parameter and tries to see how much discrimination can be obtained from it. The former approach has traditionally been followed by experimenters whose eventual aim was to build a workable speech recognizer while the latter approach has been used by people involved in speech therapy to help correct specific speech problems. An additional distinction between the approaches is that the former have tended to be much more expensive.

In the speech recognition type of display utilization, the display produces a visual image from a sound input and the viewer has to decide what utterance, out of all possible utterances, is being displayed. In the most powerful form of this display, the speech typewriter, the output would consist of the typed version of the word or words spoken. It can be argued that this is not a display but rather a full-fledged speech recognizer. In any case, we will ignore it for the present. In the less powerful forms, this type of display produces an output

image which represents some transformation of the speech input and which the viewer, possibly only after much practice, is expected to recognize.

The purpose of the present project is eventually to develop a display system which can be used as a visual feedback link for pronunciation. At the most advanced level, we might have a system which would analyze the user's utterance, compare it with some standard, and then flash on a "yes" or "no" light. However this would involve a much better knowledge of speech and the speech mechanism than is currently available. It would also provide no information about what was particularly wrong about the utterance. Thus the purpose of the present project was to eventually develop a visual display system which would present the transformed image of the user's utterance along with an image of the standard. The standard might be an idealized form generated by the display unit or it could be the version just spoken by an instructor. In either case it would be the task of the user to correct the image of his version by repronouncing it until it approached the given standard to within the appropriate tolerances.

Such a system could be used in any situation in which a person requires a visual corrective feedback path to improve his speech. One excellent example is that of people who have been deaf from a very early age. Because they are unable to hear their own voice or the voices of others, it is very difficult for them to learn correct pronunciation. A visual feedback device would be very helpful in such a situation. A second example, though not as desperately important, would be in the area of foreign language teaching in which the visual feedback could be used as a supplement to conventional language training.

In order to develop this type of display system, several steps must be taken:

1) A suitable transformation must be found to transform the spoken speech input into some format capable of being displayed.

2) Depending on the type of display chosen, tolerances must be developed so that it is possible to tell when two spoken utterances are acceptably close.

3) A suitable technique for instructing students in the use of the display must be developed since it is doubtful that any of the displays will be suitable for use without some period of instruction and practice.

The purpose of this study was to investigate various types of speech displays, to produce acceptable simulations of several of these displays using a computer-driven graphics display system, to develop some type of standardized evaluating procedure for speech displays, and apply this standard procedure to certain selected types of displays.

The remaining sections of this report can be read more or less independently. Section 2 is an elementary discussion of the characteristics of speech with an emphasis on those details which can cause trouble in speech recognition and speech display systems. Section 3 traces the history of the development of the various types of speech displays. Section 4 contains a discussion of the simulation, testing, and evaluation procedures to be used in the study.

Sections 5 and 6 contain, first a description of the various displays, and then a summary description of the computer programs used in the simulation. A more detailed description of each program, including

the listings and various test programs, can be found in Nordmann [1971].

Section 7 discusses the results of a preliminary evaluation study while section 8 summarizes the results and conclusions of the study and outlines further possible avenues of research.

Section 9 contains the list of references used in the report.

CHARACTERISTICS OF SPEECH

2.1 Problems in Speech Analysis

Speech processing devices have long been plagued with various problems which result from the characteristics of speech itself and from the effects of individual speaker differences. As Liberman, et al. [1967a] have explained, "the sounds of speech are a special and especially efficient code on the phonemic structure of language, not a cipher or alphabet". What this means is that the phonemic message being transmitted is highly restructured at the level of sound. As a result, the speech signal characteristics of a given phonemic unit vary greatly according to context.

The basic biological reason for the recoding is the fact that both the ear and the vocal articulators are slow speed devices, so that in order to deliver information at a higher rate, it is necessary to operate in parallel at both ends of the communication channel. Thus a given speech characteristic will, in general, give information about more than one phoneme and a given phoneme will be determined by more than one particular set of speech characteristics. Obviously this characteristic of speech greatly complicates any attempts at speech processing.

Bobrow and Klatt [1968] have discussed a variety of the more mundane problems involved in speech processing. Some of these problems are as follows:

- 1) The intensity range from one utterance to the next varies tremendously due to different amounts of vocal effort on the part of the speaker and the varying distance between the person speaking and the microphone.

2) The onset time of an unknown word is not a simple feature to detect reliably. This is true especially for certain initial voiceless consonants. It is also fairly difficult to separate the various phonemes which make up an utterance because the parallel operation of the speech mechanism does not produce a clear cut phoneme boundary. The most successful methods developed so far (e.g. Reddy [1966], Hughes and Hemdal [1965], Sakai and Doshita [1963], Otten [1964c], etc.) involve the establishment of certain parameters of the speech signal which are measured over extremely short periods of time. The behavior of these parameters from one time interval to the next then serves to establish whether the particular interval is the beginning of a new phoneme or a continuation of the previous one.

3) The duration of a word is highly variable. In addition, an increase in speaking rate is not accompanied by decreasing the length of time for each phoneme by the same proportional amount. For example, the time needed to pronounce stop consonants such as "p" or "b" is not as greatly affected by changes in speaking rate as is the time needed for vowels. Thus the time normalization problem is non-trivial.

4) Variations in stress and accents can greatly change the acoustical properties of the speech signal.

5) Each speaker has a different vocal cavity configuration and as a result, each speaker generates a speech signal with a different spectral configuration.

These problems, although originally discussed in the context of speech recognition, are also critical sources of variance in speech displays. In order to produce an effective display some means must be found for reducing or normalizing the effects just mentioned and

accentuating the effects which are relevant to distinguishing between different phonemes and words.

In the system being proposed this will be done by using two displays where the first display is produced by the subject and the second is presented as a standard. The task of the subject is to compare the two displays and to decide in what particulars, if any, they differ. It is hoped that most of the normalization problems can be solved by a combination of using the proper physical display and training the human observer to perform the proper pattern recognition tasks. After sufficient training the subjects should be capable of making the proper generalizations between two displays and determining the relevant points of difference and similarity.

2.2 Significant Parameters of Speech

In order to make the observers' task as easy as possible, the display should present only those speech parameters which are necessary for the recognition of the speech itself. Over the past twenty-five years a variety of research has been carried out in the search for these "significant parameters".

One of the more important features is the frequency structure of the speech wave. This structure typically peaks at three or four frequencies due to the resonating effects produced by the oral cavity during the production of speech. These peaks are called formants (Potter [1947] originally called them "hubs") and are most prominent during vowels and other voiced sounds. They are numbered beginning with the lowest frequency first.

Although the absolute frequency ranges of the various formants overlap from one speaker to another and from one utterance to another

by the same speaker (Campanella, et al [1965]), the relative position of these formants appear to be important in determining steady state sounds such as vowels (Potter [1947], Fry [1958]). In particular, it appears that the relationship of the formant frequencies of a given vowel to the formant frequencies of the other vowels spoken by the same speaker are important in the identification of that vowel (Ladefoged [1957]). Thomas [1966] has also shown that the second formant is the most important in this respect.

An even more important feature appears to be the transitions which the formants make during speech. These transitions occur as the vocal apparatus changes its configuration in order to pronounce the next phoneme in a given word. The Haskins Laboratories have done a considerable amount of work in this area by using a speech synthesis technique, in which various formant structures are converted to speech, and then checking this synthetic speech in its similarity to real speech (DeLattre et al. [1955], Harris, et al. [1958], Liberman [1957], Liberman et al. [1954], Liberman, et al. [1948], etc.). J. P. Radley [1956] has criticized this technique in that it used synthetic speech, but when he performed analyses of real speech, many of his results were similar. A summary of the cues which are useful in studying formant structure is given in Liberman, et al. [1959].

In addition to working on the transitions, Radley noted that sound bursts in the high frequency region were also important, especially in consonants such as "p", "t", and "k". Halle, et al. [1957] and Fry [1958] have also discussed this and Fry observed that it is necessary to measure the duration of the noise as well as its spectral qualities.

A different method of characterizing speech has been proposed by Roman Jakobson, Fant and Halle (see Jakobson, Fant, and Halle [1952]

and Jakobson and Halle [1956]). This method sorts out sounds using decisions based on the presence or absence of certain distinctive features such as voicing, nasalization, etc.

Table 1 gives a partial listing of some distinctive features and their values for certain phonemes. Various authors differ as to what is included in the list of distinctive features. The list in figure 1 is a composite of several different lists.

The important point as far as speech recognition is concerned is that the features can be determined independently and each has only a few possible values (usually only 2). This makes these features an ideal analysis method since the values of the various features can be determined from the speech wave without resorting to highly precise measurements.

In general the use of distinctive features has been somewhat successful in speech recognition (Hughes [1961] and Hughes and Hemdal [1965]) but has found only limited use in speech displays. This latter fact may be due to the difficulty of producing an adequate display of 8 or 10 variables. In the one example known to this author (Upton [1968]) the display was specifically designed as a supplement to normal lipreading and as a result only displayed those features which were specifically hard to see from lip movements alone.

In addition to the various types of information already mentioned, there are other types of speech parameters which might prove useful. Potter [1945] has suggested that pitch must be shown if a display is to be used for speech correction. This is certainly one of the speech functions which is most often involved in attempts to correct the poor speech habits

	o	a	e	u	ə	i	l	ŋ	ʃ	ch	k	ʒ	j	g	m	f	p	v	b	n	s	th	t	z	d	h
vocalic	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
sonorant	+	+	+	+	+	+	+	+	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-
consonantal	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
compact/diffuse	+	+	+	-	-	-		+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
grave/acute	+	+	-	+	+	-									+	+	+	+	+	-	-	-	-	-	-	-
flat/plain	+	-		+	-																					
nasal/oral								+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-
tense/lax									+	+	+	-	-	-		+	+	-	-		+	+	+	-	-	+
continuant									+	-	-	+	-	-		+	-	+	-		+	+	-	+	-	-
strident/mellow									+	+	-	+	+	-							+	-	-	+	-	-
place of articulation								B	B	B	B	B	B	B	F	F	F	F	F	M	M	M	M	M	M	M
Voicing								+	-	-	-	+	+	+	+	-	-	+	+	+	-	-	-	+	+	-

Table 1 Distinctive Features

of the deaf. Its significance in speech display applications which do not involve deaf subjects is probably not as great, although it still may be of some importance.

Chapter 3

HISTORY OF SPEECH DISPLAYS

3.1 Early Displays

The first devices which were used to make speech visible were mechanical in nature and were used for speech correction purposes. Several types were in existence in the early 1900's which utilized flames into which the subject's speech was directed by means of hollow tubes. The successive waves of dense and rarified air caused variations in the number of ions available to the flame and consequently caused the flame to flicker in a manner characteristic of the speech qualities of the subject. Abramson [1952] describes several of these devices and how they are used in speech therapy.

Characteristically, these devices were able to produce only a very gross display of the speech and about the only thing that could be determined from them was the pitch, presence of nasalization, or the relative volume of the speech. However, this is often quite helpful and due to the low cost of these devices, some of them are still in use.

Another very early type of display was an ordinary speech signal (i.e. microphone output) vs. time display. Abramson [1952] and Pronovost [1947] in their surveys on visual speech aids mention oscillographic displays but generally these displays do not give much useful information. Flowers [1916] was able to produce one of these displays in 1916 without the use of an oscilloscope by using a string galvanometer. An arc lamp projected the shadow of the galvanometer's silver-plated quartz fiber onto a perpendicular slit behind which a photographic film was moving perpendicular to the motion of the string. When a subject spoke into a microphone attached to the galvanometer, a picture of the speech signal as a function of time was produced.

There were several other types of devices which were discussed by both Abramson and Pronovost which have also been called visual speech aids. However, in many cases these devices are quite passive. Two examples are the so-called "Lite-O-Letter", a game-like device utilizing a display of transparent letters which can be lit by push buttons and the "Chromovox" (also described by Cavanagh [1951]) which involved a moving display of words and pictures to be spoken by the deaf pupil and a series of lights controlled by the teacher and used for reinforcement. Since these devices depend entirely on the skill of a speech therapist to judge the correctness of the speech sound and activate the proper indicator, they will not be considered any further here.

3.2 Spectrographic Displays

The emphasis on the more modern, electronic displays began with a Bell Telephone Laboratories project which was started early in 1941. A device for the visual translation of sound was needed in order to carry on some special studies in speech distortion which were part of the war effort. Once the needs of the military had been accomplished, however, it became possible to work on the device with the view of producing a form of "visual hearing".

The device itself was called "the sound spectrograph" and produced a three-dimensional representation of the speech signal in which time was plotted on the horizontal axis and frequency on the vertical axis with the intensity of the particular frequency component at a given time being represented by the intensity of the display at that point. Later a variety of displays were developed using three-dimensional formats with the time dimension being represented along the horizontal axis. For the remainder

of this paper this display format type will be referred to as a linear time display.

The first published reports of spectrographic linear time displays began to appear as soon as the war ended and for several years thereafter (Kopp [1946], Peterson [1954], Potter [1946], Riesz and Schott [1946] and Steinburg and French [1946]). There were actually several different types. One of the first types (Koenig, et al. [1946]) produced a permanent record by repeatedly analyzing the speech signal with a variable center frequency filter and displaying the rectified filter output on a piece of paper by means of a variable intensity stylus. Another model (Dudley and Gruenz [1946]) used a moving phosphor belt and parallel filters to display the signal in real time. Still a third (Mathes et al. [1949], Johnson [1946] used a magnetic disk and CRT system which recorded the signal and then replayed it many times at very high speed using a variable filter to give a rapid CRT display.

In 1947, Potter, Kopp and Green published the first edition of their book, Visible Speech [1947], which described the work they had done at Bell Laboratories. They had attempted to teach people to read the spectrograms they had produced much as you would read a book.

They began with a group of five young women in the fall of 1943. The instruction schedule called for two hours of group instruction and one hour of individual study each day. The following year four more young women were added to the group and also a male electrical engineer who was congenitally deaf.

The learning rate for the newcomers to the group was about 3-1/2 words per hour of study. The engineer eventually achieved a vocabulary of 800 words. The four female newcomers achieved between 100 and 300 words, but they had not practiced as long. Within the limits of their vocabulary,

the visible speech class members were able to converse by enunciating clearly and at a fairly slow rate. Potter remarked that intelligibility was roughly equivalent to a very noisy telephone connection.

Later on the original Visible Speech Translator was moved to the Detroit School for the Deaf, where Kopp and Kopp [1963a, 1963b] used it to teach speech intonation and stress to deaf children. Similar versions based on its design were fabricated at other locations as well (e. g. House, et al. [1968]). In 1965-1966 a transistorized version of the translator was produced at Bell Telephone Laboratories. Stark, et al. [1968] have reported on its use as a training aid for deaf subjects. They found that especially in the case of younger subjects, the display was of significant help but that supplemental speech instruction was also necessary.

As interest in speech spectrograms grew, various other groups designed devices for producing them. The Haskins Laboratory began speech investigations using synthetic spectrograms and a "pattern playback" device which was a "spectrograph" in reverse. Ramaswamy [1962], Harris and Waite [1963], Presti [1957, 1966] and many others developed spectrographs of varying speeds. However, they all produced the same general type of display, differing only in the way the display was produced.

3.3 Spectrographic Variations

Unfortunately there were several problems with the sound spectrograph. In addition to the poor overall quality of transmission, one of the major problems was that some of the more important features which were necessary for distinguishing between different words were not always easy to see on the display. Therefore, as time went by, various improvements were attempted.

Koenig and Ruppel [1948] describe several methods for increasing the visible dynamic range of the spectrogram. One method involved using a dot display where the density of the dots represented the intensity of the specific frequency component. Another method, which was also described by Prestigiacomo [1962], used contours to display the intensity. A third method which was further elaborated by Kersta [1948] involved reducing the spectrogram to a frequency vs. frequency magnitude plot only for specific instants of time. This allows the frequency distribution to be shown in more detail but drastically restricts the number of time intervals shown.

Another modification was one by Kock and Miller [1952] in which a differentiated version of the spectrogram was used. The display involved the differentiation of the time-amplitude pattern for different points on the spectrum. The advantage claimed for this method was that rapid changes in spectrum content, which tend to contain the most phomemic information, show up more easily.

D. E. Wood and T. L. Hweitt have described another modification [1963, 1964] in which a real time spectrograph was used to display just the peaks of the spectral cross sections. This eliminated the need for intensity modulation of the visual display. This display, as does the Kock and Miller display, emphasizes the formant frequency excursions since it is not cluttered with as much "background" data. In use as a speech analyzer this display was quite informative. However, it was not completely satisfactory in the case of stop-consonant bursts and other such signals.

3.4 Other Linear Time Displays

As more work was done with spectrograms their limitations became increasingly apparent. Although they were a good means of displaying the detailed information for an analysis of speech, they could not be read easily or quickly. As a result, several other linear

time displays were tried. These displays used the same format but processed the speech signals using different techniques in the hope that they would be easier to "read".

A display by Biddulph [1954] and an earlier one by Bennett [1953] utilized autocorrelation functions and displayed the delay parameter, τ , vs. time with the magnitude of the autocorrelation function being shown as the intensity. As it turned out, this display was actually harder to read than a spectrogram since it became very sensitive to non-critical information in the speech signal. Huggens [1954] and Stevens [1950] have each given a detailed analysis and critique of this method. Huggens shows that slight changes in pitch may cause large changes in the display. One other undesirable characteristic of the display was that it was a quadratic function of the frequency components and thus a large dominant frequency could obscure the effects of smaller amplitude frequency components.

3.5 Two-Dimensional X-Y Displays

All of the displays discussed so far have been linear time displays utilizing three display parameters. Another type of display format which has been developed involves only two dimensions in which time is generally omitted as a direct display parameter. Instead these displays use the chosen parameters as "x" and "y" inputs in a plotter (usually a CRT) which then plots the resulting point as the parameters vary with time. By using time only in this indirect sense, the originators of these x-y displays hoped to eliminate the effect on their displays of varying time duration between different utterances of the same word.

One type of x-y display which was developed utilized 90° phase shifting circuits. In this type of display the processing hardware

converted the original speech input into two output signals which were 90° out of phase with one another. Lerner [1952, 1959], Vilbig [1954], and Barton and Barton [1963] have all described displays of this type. These displays have been tested by several people but the results are inconclusive. J. E. Connor [1955] and F. E. Fabian [1955] evaluated the effectiveness of Lerner's display in speech correction and claimed that it was just as good as but no better than "conventional" speech therapy in the case of articulation disorders but of no significant help in voice improvement. However, in a later preliminary study, Pronovost [1964] felt that this display showed some promise in improving the articulatory proficiency of deaf children. Unfortunately, a subsequent study (Pronovost, et al. [1968]) was unable to produce more definite results.

Pyron and Williamson [1964] gave a critique of Barton and Barton's apparatus and indicated what they thought was the general problem with all such techniques, namely that they work best on continuous sounds (i.e. vowels and nasal consonants) and are very poor on transitions (i.e. consonants) which carry a high proportion of the speech information.

A different type of x-y display has been developed in Switzerland by Dreyfus-Graf [1946, 1948, 1950a, 1950b]. This display uses a system of filters and differentiators to produce pulses which control the movement of an ink pen. The author claimed that the resulting squiggles, which do appear fairly consistent for sustained vowels, could be used as a phonetic shorthand. However, as far as this author knows, there has been no report on the use of this device with a normal speech input.

Another x-y display using a CRT has been reported by Plomp, Pols and Van de Geer [1967]. They analyzed 15 Dutch vowels by using a bank of 18 filters to process the speech signals and studying the differences between the vowel spectra. The resulting dimensional analysis yielded four dimensions which accounted for 96.4% of the total variance once the between-subject variance had been allowed for. The authors suggested using plots of the first dimension vs. the second as an aid for the deaf. An oscilloscope display for the vowels has been produced but work is only beginning on the consonants. This method was suggested as an alternative to a type of display in which the frequencies of the first and second formants for various vowels are plotted as points or regions on a two-dimensional graph (see for example Davis [1952], Foulkes [1961], Hughes [1965] or Majewski [1967]). Although this type of a representation is very appropriate for vowel sounds it has not met with much success in the representation of consonants. It remains to be seen if Plomp, et al. will be able to apply their technique to the consonants.

Cohen [1968] has described an x-y display developed by Arthur D. Little, Inc., which was made from a converted TV set. It used a type of frequency analysis somewhat similar to the cepstrum analysis technique (Noll [1964, 1967]) in which the log of the output of a spectral analysis is subjected to another "spectral analysis" to determine "shape" characteristics of the original spectrum. The ADL display makes use of 10 filter channels and by the use of various weighting factors resolves their outputs into sine and cosine components of the frequency spectrum envelope. These two components are then plotted as the x and y coordinates of the display. The net result is somewhat analogous to a two-formant display

but the problem of formant identification is avoided. The device is currently undergoing evaluation for use by deaf people for speech improvement.

3.6 Zero-crossing Displays

In addition to classifying visual displays according to their physical format, they can be separated according to the type of processing used on the speech signal. Thus we have already discussed spectrographic, correlation and phase splitting displays, among others. Another very common type of processing is the extraction of zero-crossing information. One of the reasons this type of processing is so popular is that it can be easily performed using a high gain amplifier and clipping circuit, and is thus cheaply implemented.

One linear time display version of a zero-crossing display was developed by Chang, et al. [1951b] and further developed by Sakai and Inoue [1960]. It was called an "intervalgram". This display used the time intervals between zero-crossings or between zero-slopes (i.e. zero-crossings of the differentiated signal) as a parameter to be plotted against time. The display produced a dot for each interval between zero-crossings where the horizontal position of the dot was determined by the relative time position of the interval and its vertical position by the frequency of the sinusoidal signal which would have produced an equivalent interval between zero-crossings. The result is a halftone display consisting of dots which look somewhat similar to a spectrogram.

C. C. Bridges [1964] has produced a more simple linear time zero-crossing display by simply plotting the zero-crossing rate as a function of time on an oscilloscope.

The main justification for using these parameters was the finding by Licklider and Pollack [1948], Licklider [1959], and others,

that highly clipped speech signals, and highly clipped differentiated speech signals were still quite intelligible to the human ear. Thus, since these clipped signals contain only interval information about zero-crossings or zero-slopes, a display of this information should contain all the essential information of speech. In addition, of course, these parameters were much easier to obtain than spectrograms or correlation patterns. However, the authors were unable to show that intervalgrams were any easier to read although Sakai and Doshita [1963, 1968] did use this technique for speech analysis and recognition.

Pyron and Williamson [1965] have developed an x-y display utilizing zero-crossing information in which they extracted the amplitude envelope of the speech signal as well as the rate of zero-crossings and the rate of zero-slopes. They experimented with plots of amplitude vs. zero-crossings, zero-crossings vs. zero-slopes, and amplitude vs. zero-slopes, but since they discovered that the latter gave consistently clearer and more characteristic patterns, most of their results are concerned with that form. As the authors noted in their report, Chang [1951a] has provided a theoretical analysis and experimental evidence to show that, in speech signals with a pronounced formant structure, the rate of zero-crossings corresponds to the first speech formant while the rate of zero-slopes corresponds to the second speech formant. Thus, their display is analogous to an amplitude envelope vs. second formant x-y display.

Ewing and Taylor [1969] have duplicated Pyron and Williamson's display and have attempted to improve upon their results. They initially worked with a zero-crossing vs. zero-slope type of display with the eventual aim of generating patterns which could be recognized by

computer. They also tried adding a time sweep to both axes which gave the display a diagonal rise across the face of the CRT. However, they felt their most promising version was one in which the difference between the zero-crossing and zero-slope signals was plotted vs. time. In this case they still did not get the desired results but they felt that this was due to poor comparison methods during the recognition phase of their procedure.

3.7 Pitch Extracting Displays

Another type of processing used in producing speech displays is pitch extraction. As early as the 1930's, Coyne [1938a, 1938b] and Timberlake [1938] reported on a voice pitch indicator using 14 to 20 mechanical band-pass filters (i.e. tuning forks) with lamps which indicated the pitch frequency. Its use in South African schools for the deaf has shown good results for younger subjects but negative results for older subjects with settled voice habits.

Dolansky [1955] has described a pitch extracting device based on a time domain analysis. The descendants of this device have been used to produce displays which have been used in several experiments. These displays are linear time displays but only use two dimensions. Time is on the horizontal axis with the position on the vertical axis indicating the pitch period of the incoming speech. The intensity of the display is turned off when no voicing is present, but other than this, is independent of the speech input.

F. Anderson [1960] has used a version of Dolansky's pitch extractor utilizing a revolving CRT with a view panel, cut so that only a portion is displayed on the vertical axis against a continuous horizontal time base. The CRT uses a long-persistence phosphor so that the display

can be seen for five seconds. The display was used with a group of eight children from ages 8 to 12, with hearing losses of 60 db or more. It appears to have been somewhat useful although the author did not go into detail about it.

The group headed by Dolansky at Northeastern University continued to work on pitch displays (Dolansky, et al. [1965], Dolansky and Phillips [1966], and Phillips, et al. [1968]). They performed several studies using deaf children as subjects as well as normal hearing university students. The results indicated that the display was of some use in teaching deaf children and that it was possible to use the display as a visual feedback indicator for speech pitch.

A variety of other researchers have developed pitch extraction displays (Gruenz and Schott [1949], Plant [1960], Martony [1968], and others). In addition, several displays have been made which incorporated a pitch display along with some other type. Stark's spectrographic display [1968], mentioned earlier, uses pitch and amplitude as well as spectrographic information. Pickett and Constan [1968] describe a multi-display device developed at the Hearing and Speech Center of Gallaudet College which in addition to being able to produce a pitch display, could also generate vowel spectrum indications, intensity vs. pitch displays, and intensity contours.

3.8 Miscellaneous Formats

In addition to the linear time and x-y display formats, there has been a variety of other types of attempts. D. E. Williams [1967] has designed a light bulb display which consists of a matrix of lights and an electronic circuit to drive it, which frequency analyses an utterance into 10 frequency regions and displays the results in a bar graph form. There

was also a second display which indicated the relative length of time each frequency component was above a certain threshold. However, the display appears to be valid only for sustained sounds like vowels and even in these cases varies tremendously with such irrelevant variables as distance from the microphone, speech rate, etc.

Hubert W. Upton has developed a wearable eyeglass speechreading aid (Upton [1968], Picket [1969], Risberg [1969]) which detects voicing, friction, stops, etc. Miniature lights imbedded in the eyeglasses glow whenever the corresponding speech feature is present. The device was specifically designed as an aid to lipreading and therefore the speech features which were chosen were those not visible on a speaker's lips. The designer noted that although the analyzing functions did not work perfectly, the device still gave a significant amount of information not obtainable by lip reading alone.

In addition to these displays, there are several other types which have been in use by speech therapists but which do not fit neatly into any of the categories mentioned so far. Risberg [1968] discusses a variety of these devices which he helped to develop, including various types of indicators for fricatives, s-sounds, intonation, rhythm, and nasalization. Some of these displays might be called linear time displays, but others involve simply meters or lights which turn on when a given threshold has been reached for the quantity being measured. The primary principle was to minimize the number of functions displayed by a single device. This was done both to decrease the cost and to isolate the speech feature to be controlled.

3.9 The Use of Speech Displays

Although a wide range of speech displays have been developed over the past twenty years, there has been a reluctance on the part of speech

therapists to make widespread use of them. The reasons for this have been briefly mentioned above and center on the cost of the devices and the pedagogical problems which they produce. From a cost standpoint, it is easy to see that the more complicated (and thus more costly) displays would badly distort the small budgets of most schools for the deaf. The pedagogical resistance might be a little harder to justify. However, although it may be true that some of the resistance is simply a result of innate conservatism on the part of teachers of the deaf, it is also true that very little testing has been performed on the effectiveness of the various display types. Thus the fears of these teachers toward using untested techniques on children whose futures may depend on them are somewhat justified.

More recently, however, the situation has been changing. A variety of small experiments have been performed to determine the feasibility of using particular displays as a visual feedback link to replace the auditory feedback link which has been destroyed in deaf people. The primary goal has been to use some type of visual display to indicate to the deaf subject how correct or incorrect his pronunciation actually is. In general, these studies have been promising for younger subjects, though not as successful for older subjects.

The tests themselves have mostly been performed by specialists in the area of speech training and have involved the simpler types of displays. Cost is the obvious reason for this latter fact. This same fact also makes it very difficult for any one group to build and test more than one or two displays at the same time. As a result there has been very little work done in developing general testing techniques which could be applied by a single group to a wide variety of displays in order

to determine the relative effectiveness of the different types. Happily this trend appears to be reversing as can be seen by the previously mentioned development of systems which can produce more than one display type.

Although many groups have been able to use speech displays as feedback aids in speech correction for the deaf, the original goal of the Bell Laboratory group, i.e. actually reading the display, has yet to be achieved. It has in fact been suggested by A. M. Liberman, et al. [1967a] of the Haskins Laboratory, that we may never be able to perform this type of direct conversion. This is so, they maintain, because there is no simple one-to-one correspondence between the characteristics of the speech signal and the phonemes which it represents (Liberman, et al. [1967b]). Since the speech signal is basically a complex code as opposed to a simple cipher, the phonemic message being transmitted is highly restructured at the level of sound. As a result, the speech signal characteristics of a given phonemic unit vary greatly according to context.

The basic biological reason for the recoding is the fact that both the ear and the vocal articulators are slow speed devices, so that in order to deliver information at a higher rate, it is necessary to operate in parallel at both ends of the communication channel. Thus a given speech characteristic will, in general, give information about more than one phoneme and a given phoneme will be determined by more than one set of speech characteristics.

The key point to their argument against the readability of visible speech, however, is their statement that although a decoder of such signals obviously exists, it appears to be unalterably linked to the

auditory sensory system. Thus although it might be possible to create displays which emphasize the important key features of the speech signal, it does not appear possible to produce a display which would allow the viewer to unconsciously decode the signal into phonemes. It should be noted that this key point, by the authors' admission, appears to be true only because in 20 years of experience nobody has been able to learn to visually decode spectrograms without a great deal of conscious mental effort.

Recently, however, Lenneberg [1967] has discussed the effect of age and development on the learning of a language. According to a variety of experiments it appears that the development of speech is impossible once a human has reached approximately the age of puberty. Before this time, humans are capable of learning language even if large portions of the brain which are normally connected with this process are destroyed by disease or accident. The brain seems to be very plastic at this age and highly adaptable.

As a result, it may be possible that the proposition put forth by the Haskins group will only hold for adults since their brains have already "frozen" into a permanent state. This could also explain why younger subjects seem to get the most help from feedback type displays. It would be interesting to try to teach a deaf child to read visible speech since in this case the child's brain might actually be able to adapt itself to decoding the visible input.

Be this as it may, if we grant the fact that the human eye cannot be trained to become an automatic speech decoder (at least once the subject passes a certain age) then the task of using a visual display as a speech feedback mechanism for adults can be looked at from

two positions. If the feedback device actually performs the decoding before presenting the visual display, then it becomes in effect, a speech recognizer. This is precisely what the last 20 years of speech research has been trying to achieve but without too much success. In addition, it would not be very helpful in the present task since it would not be giving the information which a poor speaker needs to correct his pronunciation.

We can, on the other hand, ignore the absolute decoding problem and instead concentrate on displaying the most relevant speech parameters in a concise manner. In this case the observer would not necessarily be able to recognize the words merely from the display. However, if the proper parameters are displayed in an easily discerned manner, it should be possible for the observer to detect the differences between his pronunciation and a comparison display of the same speech pronounced properly. This is the eventual goal which has been set up for this project.

Chapter 4

PROPOSED STUDY

The eventual aim of this research is to develop a computer driven display system which can be used as a visual feedback link to correct mispronunciations by people who are deaf or in other situations, such as language training, where correcture feedback in pronunciation may be desirable. The envisioned system would present two displays to the user, one of the word as it is supposed to be pronounced and one as it is pronounced by the user. His task will be to determine if they are acceptably close (this may be possible only after a certain amount of instruction and practice) and if not, to determine which parts are in error and change his pronunciation accordingly.

The more immediate goal of this particular study has been the development of a generalized computer simulated display system. This system has been built so that it can utilize a variety of speech processing techniques and easily produce a wide range of speech display types. In addition several of these displays can be compared with one another to determine which of them is most effective in terms of presentation of relevant variables and ease of training in their use.

The speech display simulation system has been implemented on the CDC 1604 installation at the Coordinated Science Laboratory at the University of Illinois. This system contains a high-resolution variable intensity CRT display equipped with facilities for taking both still and moving pictures.

The main advantage of using such a system to generate speech displays is the flexibility inherent in a computer simulator. Using this type of system it is extremely easy, once the basic processing programs

have been written, to modify displays and to create new ones. There are no time consuming hardware modifications to be made. Of course the main disadvantage is the cost of the computer system. Once a suitable display design has been found, however, a hardware version can be fabricated. Alternatively, a time sharing educational system such as the PLATO system might be used to allow access to a large scale computer at a minimum cost. If low cost display units such as the plasma panel (see Bitzer, et al [1966] or Willson[1966]) become readily available and are capable of producing the type of displays needed, this latter implementation might be an inexpensive way of providing a variety of display types to the various institutions needing them.

4.1 Outline of the Study

The development of this study was organized into the following steps:

- 1) The development of a basic subsystem for inputting speech signals into the computer. Because of the slowness of the CDC 1604, it was not possible to run the complete speech display simulation system in real time. As a result the speech input subsystem has been oriented around the tape units as a storage medium. The I/O programs were used to read in data from an audio tape recorder attached to an A to D converter and to write out the data in a packed format on magnetic tape. This data was edited by the operator by means of various data manipulation programs. Eventually the desired data was copied to a new tape complete with header blocks describing the data. This edited data tape was then used as the input to the processing routines.

- 2) The development of various speech processing routines general enough to be used by a variety of display types. These include

such routines as peak detectors, zero crossing detectors, rectifiers, fast Fourier transforms, digital filters, etc.

3) The development of the speech display routines. These routines make use of the speech processing routines plus other types of general routines to produce specific displays. The types of displays are explained in detail in Section 5. Some were designed from descriptions in the literature (see Section 3) and others were developed more or less independently. As new speech processing routines were found to be necessary, they were developed and added to the programs written in stage 2.

4) The production of display photographs for use in experimental comparison tests. A limited number of words were picked and recordings of these words correctly spoken by several people were made. After being converted to digital tape, these recordings were processed by the various routines to produce the desired displays. These displays were photographed using a polaroid camera and each resulting picture was rephotographed to produce a 35 mm. slide, which could be shown to subjects by means of either a slide viewer or a projector.

5) Finally, two types of tests were conducted on each of several types of displays to determine their relative effectiveness in displaying speech. A preliminary test was conducted for the availability of the proper information for word discrimination. The preliminary test was a type of concept attainment experiment in which the subjects must try to identify each word from its display. The point of the preliminary test was to determine if a given display type presents the proper information for word identification. In other words, is the transformation appropriate? Since it is fairly well established that this type of concept identification task is a hard (if not impossible) task in the general

speech display case, this test was made using a limited number of words.

A final test to determine the displays' usefulness in a comparison situation such as would exist in the eventual system was also conducted. In this test the subjects were presented with pairs of photographs which represented two different utterances as depicted by one of the display types. The two utterances could be the same word spoken by two different people, different words which sound similar, or a correctly and incorrectly pronounced version of the same word. The subject's task was to determine if the two displays represented the same word. After his response he was told the correct answer. As a further test the subject was occasionally asked to indicate points of similarity or difference. Then on the basis of the number and type of errors made on each display, a comparison between the various display methods was made.

With the completion of the comparison tests the scope of the present study ended. There are still other problems. In particular the question arises that even if the subject can correctly detect a difference between two displays, he may not know how to change his pronunciation to make the display of his version of the utterance more like the standard. However, in order to test out this problem a real-time display is essential. Therefore for the time being, this problem will be postponed.

In conclusion the goal of this study was to develop several types of visual speech displays and then perform comparison tests on them to determine their relative and absolute suitability for use as visual speech feedback devices.

4.2 Theoretical Significance of the Comparison Tests

As was previously mentioned, the theoretical basis for the comparison tests used in this study comes from that part of the psychological

literature dealing with cognitive processes, which has come to be called "concept identification" or "concept formation". The tests themselves involve the establishment by the subject of various response categories, i.e. the words, based on generalized concepts which must be developed by looking at the various instances of these categories as depicted by the particular display type being tested. In order to do this the subject must select those attributes from the display instances which are most relevant to the discrimination process and determine how these attributes indicate the proper response categories.

Over the past few years there has been a great deal of discussion in the literature of concept identification about the exact method used by subjects in the development of concepts in this type of situation. Restle [1962], Bruner, Goodnow and Austin [1962] and Haygood and Bourne [1965] all discuss various types of strategies for selecting and testing hypotheses about the cues which will lead to a correct classification. Haygood and Bourne [1965] break the process down into two problems: finding the attributes of the various instances which are important in determining the concept(s) and finding the rules involved in combining the values of these attributes. The attributes may vary in their obviousness and the rules may be either simple, i.e. merely the presence of a particular value of the attribute, or complex, i.e. some logical relation between several attributes.

Bower and Trabasso [1963], in discussing two category problems (i.e. the concept is simply the presence or absence of a particular value of one of the attributes), develop an expression for the probability that the subject will focus attention on the relevant attribute, namely:

$$r = \frac{W_a}{W_a + W_i},$$

where W_a is the attention value of the relevant attribute summarizing all of the factors determining the subject's selection of it for testing and W_i is the sum of these values for the irrelevant attributes. In a more complicated situation, such as the present case of speech displays, there are other factors to be considered as well.

In the first place, there may be redundant attributes which would help to establish the response categories. These may be wholly redundant or they may be only partially redundant and thus only help in some of the cases. Secondly the rules involved in combining the attributes are probably more complex than the simple presence or absence of a particular value of an attribute. Some of this complexity will be due to the inherent complexity of the speech code, and some of it will be due to the partial redundancy of some of the attributes. Finally, the fact that we are working in a multiple-response category situation will increase the complexity of any such formulation.

As a result of all of this, it would be very difficult to develop any kind of precise mathematical formulation for the probability of achieving concept discrimination in the present case, and in fact this is not really necessary. All we actually need are a few qualitative predictions.

Basically, the preliminary test is meant to be a concept formation situation in which the subject must learn to identify words based on the cues being presented by the display type being tested. It is hypothesized that the speed with which the subject attains the "concepts" of the words as represented by that particular display is directly related to the probability of concept attainment after a number of trials. This in turn

is hypothesized to be related to the number and effectiyeness of the relevant cues and inversely related to the number of irrelevant cues presented by the display. If the words selected for display are sufficiently typical of the normal speech sounds encountered in spoken language, and if several speakers are used to get a typical set of speaker variations, then, provided that there is a difference in the effectiveness of the various displays, the results should be significant. By measuring the length of time it takes to achieve a given criterion of performance on a particular display, we should obtain an indication of the relevance of that display type to the problem of word identification.

The purpose of the second test was to determine, for each display, the type of variations of the words which can be accepted as unimportant. Since the second test takes place after the subject has gone through the first test phases, the subject will have become somewhat proficient (hopefully) at understanding the display. Thus this test is akin to a concept discrimination task in which the subject is trained to make finer and finer distinctions.

Chapter 5

DISPLAY DESCRIPTIONS

The purpose of this section is to give a detailed description of the various types of displays which have been produced by the Speech Display system. Each display will be described separately in general terms along with the different variants which are possible. Photographs of these various displays will also be given.

Before describing the speech display types themselves, it will be desirable to describe the two main display packages which these speech displays use, the variable-intensity TV scan display and the continuous line display.

5.1 Variable-Intensity TV Scan Display

This display program package takes a two-dimensional array of intensity points and produces a continuously varying intensity display. The programs interpolate between the points in the array in both dimensions and set up a TV scan display buffer which is plotted and photographed by the system display routines. There are a variety of display choices which can be specified by the user:

- 1) The number of points to be interpolated between the array entries in both the horizontal and vertical directions.
- 2) The distance between points which are plotted by the system display routines (this will affect the "grain" of the resulting display).
- 3) The position relative to the left-hand side of the display at which the actual data will begin to be displayed. (This allows a given speech display to be centered.)
- 4) The minimum intensity below which the data will not be displayed. (This helps to eliminate low-intensity clutter which takes time to display but which adds no real information.)

5.2 Continuous Line Display

This display package produces a continuous line display using either an x vs. y type data format or a format in which one data array is plotted sequentially in the horizontal direction. The main display options are the maximum x and y values and the type of display.

5.3 Spectrogram

At the present time the spectrographic display is the most versatile display in the sense that it can be varied in the most number of ways. As described in section 3, it is a linear time display in which frequency is plotted along the vertical axis and time along the horizontal axis. The intensity of the display at any given point is proportional to the magnitude of the particular frequency component at the time represented by that point.

The actual frequency analysis is done using a Fast Fourier Transform program initially written by Gary Horlick of the Coordinated Science Laboratory and subsequently modified by the author. The algorithm is a variant of the original Cooley-Tukey algorithm (see for example, Cooley and Tukey [1965], Gentleman and Sande [1966], Cochran, et al [1967], or Brigham and Morrow [1967]). More recently, Alan Oppenheim [1970] has presented a very good article on the use of the FFT in producing spectrograms.

Since the FFT is a discrete transform, the output frequency magnitudes are, in effect, samples of the frequency spectrum of the data being analyzed. The spacing of these frequency samples is determined by the fundamental frequency of the time period being analyzed, and this in turn is determined by the number of samples being processed. Thus it is possible to decrease the spacing between frequency samples by increasing the number of time samples processed. This produces a more detailed frequency analysis but only at the cost of having a larger time slice.

This effectively means that although you have gained more information about the frequency analysis, you are less sure about the position in time to which it applies.

In addition to the time-frequency tradeoff, it is also possible to adjust the number of frequency components to be displayed and thereby vary the total frequency spread of the display.

Once the frequency components have been calculated for each time slice in the display, a linear normalization is performed on the data so that the intensity values will be within the range of values used by the CRT. The value given to the maximum component in the display can be adjusted to be greater than the maximum intensity which can be displayed by the CRT. Since any intensity values greater than the maximum displayable value are truncated to the maximum intensity value by the display routines, this allows the user to specify a value range over which he desires truncation of the intensity values. This feature is valuable because in any given spectrogram there are always a few points which are way out of line with the rest of the values and by truncating these points, the remaining points can be given a greater spread of values.

A second form of contrast enhancement can be used, namely high frequency emphasis. This simply involves multiplying each frequency component in every time slice by a factor greater than or equal to 1 and having the factor increase as the frequency of the component increases. In the actual program the emphasis begins at around 2000 hz. and the user in effect controls the rate of increase in the multiplicative factor.

In addition to these options, the spectrographic display can make use of the various options available in the variable intensity display package. Figures 1 through 3 show various examples of spectrographic displays using various sets of parameters.

a) no emphasis
no truncation

b) medium emphasis
no truncation

c) high emphasis
no truncation

d) no emphasis
medium truncation

e) medium emphasis
medium truncation

f) high emphasis
medium truncation

g) no emphasis
high truncation

h) medium emphasis
high truncation

i) high emphasis
high truncation

Figure 1 Effect of Variations in High Frequency Emphasis
and Intensity Truncation Using the Word "Shod"

nsamt = 64

nsamt = 128

nsamt = 258

nsamt = 512

nsamt = 1024

Figure 2 Effect of Variations in Time Slice Size

Speaker a
"shod"

Speaker a
"vile"

Speaker b
"said"

Speaker c
"said"

Speaker b
"ted"

Speaker d
"dame"

Figure 3 Examples of the Spectrographic Display
with Nominal Parameter Values

5.4 Formant Extracting Display

The formant extracting display is similar in format to the spectrographic display. However, in this type of display the formants are extracted from the display data and all other display data in the frequency regions of the formants is suppressed. This allows the formant movements to be seen more clearly and at the same time retains the high frequency fricative information.

The formant extracting process essentially takes the frequency analysis of each time slice and finds its major peaks. This involves utilizing a peak-picking routine twice (see figures 4 and 5). The first pass over the frequency analysis data obtains the minor peaks which represent the various harmonics of the fundamental pitch frequency. The second pass of this data will obtain the peaks which can be considered to be formant candidates.

The four largest formant candidates are then selected and analyzed. If the smallest candidate is less than half the size of the next smallest, it is eliminated. Any candidates over 4000 cps are also eliminated since it is unlikely that a true formant would appear in that frequency region.

Once the unlikely candidates are eliminated, the frequency analysis in the region covered by the remaining formants is erased and replaced by the magnitudes of the formants at their corresponding frequencies. The results of this type of analysis are shown in figure 6.

It should be noted that this algorithm never determines which formant is the first, which is the second, etc. This is a non-trivial problem since some of the "formants" selected may still occasionally be noise and quite often "real" formants will drop out or merge for several time slices. The only way to effectively determine the actual number

Initial Spectrum Analysis
of Time Slice

First Pass Selects All
Peaks Due to Pitch Harmonics

Second Pass Selects
Potential Formants

Final Formant Selection

Figure 4 Effect of the Peak-Picking Process
on the Spectrum Analysis of a Single Time Slice

Initial Spectrograph

Display of Peaks
After First Pass

Display of Peaks
After Second Pass

Final Formant Display

Figure 5 Effect of the Peak-Picking Process
on the Full Spectrographic Analysis of the Word "Dead"

Speaker a
"shod"

Speaker a
"vile"

Speaker b
"said"

Speaker c
"said"

Speaker b
"ted"

Speaker d
"dame"

Figure 6 Examples of the Formant Extracting Display

associated with the formant would be to keep a record of the movements over time and on the basis of this record determine which peaks in a given time slice correspond to each formant.

5.5 Zero-Crossing Display

The zero-crossing display is a linear time display in which the frequency equivalent to the zero-crossing rate is plotted on the vertical axis and time on the horizontal axis. The speech input is fed to four digital filters, the outputs of which are then analyzed to determine their zero-crossing rates. A single point is plotted for each filter output, the magnitude of the point being proportional to the magnitude of the output of the corresponding filter.

The frequency regions have been chosen so as to approximate the regions covered by the first, second, and third formant, with the fourth region being a high frequency region for fricatives or other noise-like sounds.

Examples of this type of zero-crossing display are shown in figure 7.

5.6 Zero-Crossing vs. Amplitude Envelope

This display is a simulation of the display described by Pyron and Williamson [1965]. There are actually two variants, one using the zero-crossing rate of the original speech signal, Z_1 , and the other using the zero-crossing rate of the derivative of the speech signal Z_2 . (This latter signal can also be thought of as the "zero slope" or maximum-minimum rate). One of these two signals is plotted against the amplitude envelope of the speech signal to produce an x-y type speech display.

A block diagram showing the production of these two display variants is shown in figure 8. Note that in producing the y input,

Speaker a
"shod"

Speaker a
"vile"

Speaker b
"said"

Speaker c
"said"

Speaker b
"ted"

Speaker d
"dame"

Figure 7 Examples of the Zero-Crossing Display

Figure 8 Block Diagram for Z_1 and Z_2 vs. Amplitude Envelope Display

i.e., the Z_1 or Z_2 signal, the differentiator is omitted if a Z_1 signal is to be used. The threshold detector is used as a blocking device to inhibit the display when no signal is present.

The display itself has been adjusted so that the zero-crossing rate axis varies from 0 to 6000 zero-crossings per second (equivalent to 0 to 3000 cps.) as was described in Pyron and Williamson [1965]. However, there is no 3KC upper limit bandlimiting as was true in their case. Thus occasionally zero-crossing rates greater than a 3KC equivalent may be encountered. In this case the display routine truncates the display in the vertical direction. The horizontal axis has been adjusted so that the maximum amplitude coming in on the A to D converter will cause a full-scale deflection.

Examples of the z_1 and z_2 vs. amplitude envelope displays are given in Figures 9 and 10, respectively.

Speaker a
"shod"

Speaker a
"vile"

Speaker b
"said"

Speaker c
"said"

Speaker b
"ted"

Speaker d
"dame"

Figure 9 Examples of the Z_1 vs. Amplitude Envelope Display

Speaker a
"shod"

Speaker a
"vile"

Speaker b
"said"

Speaker c
"said"

Speaker b
"ted"

Speaker d
"dame"

Figure 10 Examples of the Z_2 vs. Amplitude Envelope
Display

Chapter 6

SPEECH DISPLAY SIMULATION SYSTEM

The Speech Display Simulation can be divided into four main areas: the common data base, the command processor, the speech display routines, and the various subprocessing routines.

6.1 The Common Data Base

The common data base consists of the input speech data buffer, BUFF, the output display data buffer, FINT, the CRT display command buffers, ISCOPE and ISCOPE1, and all of the constants and variables used to control these buffers. These buffers and variables are all kept in COMMON storage. The problem of keeping the COMMON declaration in each subroutine identical is handled by means of the CSL FORTRAN title feature. This extension of the FORTRAN language allows the programmer to specify FORTRAN statements which will then appear in every program in which the statement TITLE* appears. Any type of valid FORTRAN statement can be put in the title and thus the whole common data base need only be written down once.

The common data base has several key features. Since the CDC 1604 was not fast enough to process speech input in real time, it was necessary to use digital tape for storing the input speech data. As a result, it became unnecessary to provide a full-sized buffer to contain a complete speech utterance. Instead, the floating point buffer, BUFF, is used to contain only that portion of the data which is of current interest.

As can be seen in figure 11, there are two corresponding pointers for the data tape and the buffer, BUFF. ISAMP is the main data pointer and selects the initial sample of a set of data points

DATA TAPE (packed integer format):

ISAMP

BUFFER (unpacked floating point format):

ISAMPB

Example: ISAMP = 54627
ISAMPB = 1627

Figure 11 Relationship Between ISAMP and ISAMPB

from the complete set of data (consisting of many speech utterances) on the data tape. Its value may range up to around 900000, since this is the approximate number of packed sample points which can be written on a single tape. ISAMPB corresponds to ISAMP in that it points to the same data as ISAMP but it refers to the data as it happens to be currently loaded in BUFF. Thus ISAMPB only varies from 0 to the maximum length of BUFF (currently 3000 words).

The display generating routines are free to move ISAMP up and down the data tape whenever they wish. Before they utilize this new data position, however, they must call the subroutine ADJUS2. This subroutine checks BUFF, and if the data corresponding to the new position of ISAMP is not currently in BUFF, it moves the tape forward or backward until it can load BUFF with the proper data and converts it to floating point. Once BUFF is made to contain the desired data, ADJUS2 sets ISAMPB so that it can be used as an index for BUFF to obtain the desired data. It is this pointer that the speech processing programs use to obtain the speech data.

The second feature of the common data base involves the FINT array. This array is basically a two-dimensional array containing intensity values with its dimensions corresponding to frequency vs. time. However, it was felt that it would be much more convenient to be able to vary the relative maximum sizes of these two dimensions even while the total length of the array remains fixed. This is especially nice for short speech samples in which it is desired to have a spectrographic analysis with a very small increment between frequencies, since in this case the maximum index for the frequency dimension must be increased. Unfortunately FORTRAN has no provision for dynamically

assigning array dimensions. Therefore it was decided to require each program using FINT to calculate its own subscripts using a frequency maximum index, IFMAX, which could be dynamically chosen by the operator. At first this seemed like a lot of extra work but the technique is relatively straightforward and in many cases it resulted in a considerable increase in speed due to the lamentably inefficient calculations used by CSL FORTRAN to calculate subscripts. This was especially true in loops since the compiler makes no optimizing attempts.

6.2 The Command Processor

The command processor is the heart of the interactive communication with the system. It gives the operator the ability to change the values of the system constants and variables and to call the various display routines. In addition, he can dump out the contents of the various arrays and variables. The command processor includes the main program and the subroutines directly called by it, namely INPTCM, which reads each command with its parameters and the various command identifying subroutines, which determine the command and perform the requested operations. At the present time, INPTCM accepts only fixed format commands. However, it is hoped that it will eventually be possible to expand it to a free format subroutine.

The command identifying operations have been kept as general as possible. The commands are grouped together according to function into subroutines. Each subroutine has the task of identifying those commands associated with it and then executing them. Since the subroutines are independent of one another it is relatively easy to expand the command set simply by adding commands to the relevant subroutine or writing a completely new subroutine and adding a call to it in the main program.

The conventions for intercommunication are relatively simple and yet allow a high degree of flexibility. Each subroutine accepts as parameters a character variable containing the command and as many of the input parameters read in by INPTCM as may be necessary. If the subroutine determines that the command is not one for which it is responsible, it simply returns. If the command is one of the subset of commands which it can execute, it performs the required operations. Then before returning, it sets the command variable to zero to indicate to the main program that the command was executed. Thus after calling all of the command identifying subroutines, the main program merely needs to check the command variable for zero to see if it was executed. If it is not, then the main program types out a message saying that the command was not recognized.

Note that this technique presents a wealth of opportunities. For example, a command identifying program, as part of its command execution step, could load the command variable with a new command instead of loading it with zero. This command could then be executed by some subsequent command identifying program. This in fact has been done in the present system. To extend the idea even more, the command variable could be generalized to a push down stack. Then you could have complex commands which actually represent a series of simpler commands. The execution of the complex command would consist of expanding it into the simpler series of commands and pushing these into the stack. The main program would pop the stack each time a command was completed and then repeat the identification and execution process for the newly exposed command at the top of the stack. The main program would only return when the stack was empty. The key point to note (and the one which illustrates the general philosophy of the system) is that

this stacking process could be added without modifying the programs which already exist.

Some of the commands which can be executed by the system are given in Table 2. In addition to being able to run the various display programs and diagnostic routines and to manipulate the data tapes, the command system allows the operator to change many of the system variables. This allows him to easily modify the various displays. It also causes a certain number of problems due to the manner in which some of the system variables and constants interact. An example of this problem occurs in the spectrographic display, where the number of samples to be processed per time slice fixes the interval between frequency coefficients and vice versa.

The solution to this problem was to allow the user to set certain parameters independently and then have the system calculate the effect of these choices on the other dependent parameters and print them out (this operation is performed by the FINI subroutine). Thus, for example, the operator can choose the desired number of data samples he wants processed per time slice in the spectrographic display and the system will respond by indicating the frequency increment between coefficients and the total frequency range which will be displayed given the current value of IFREQ.

6.3 The Speech Display Routines

The speech display routines consist of the programs used to simulate the various speech displays. These programs manipulate the common data base using the various subprocessing routines to produce the displays desired.

Command	Subroutine Which Executes Command	Command Operation
BEGN	TAPCOM	Rewind data tape & initialize system
BUFF	DIAGNG	Print out buffer contents
C	DIAGNG	Next input will be a comment
COPY	DATGCL	Copy data tape
DISP	DIAGNG	Display buffer contents on CRT
F	TAPCOM	Short form of FOWD = 1000
FINIS	PROSCL & DATGCL	Calculate dependent variables & turn off CRT
FIND	TAPCOM	Search data tape for specified speech words
FORME	PROSCL	Call FORMEX display routine
FOWD	TAPCOM	Move data tape forward NVAL samples
HEADT	TAPCOM	Process header block
HIEMP	PROSCL	Add high frequency emphasis to display data
INITI	PROSCL & DATGCL	Initialize system variables
INTAP	DIAGNG	Assign input command medium
IWIDE	TAPCOM	Assign window size for data tape display
LOCA	TAPCOM	Print out value of data pointer
MOVE	TAPCOM	Move data pointer to NVAL
NORMF	PROSCL	Normalize display data
OBTAI	DATGCL	Use A to D convertor to obtain speech data
PHOTO	PROSCL & DATGCL	Take picture of last display
PYRON	PROSCL	Call PYRON display routine
READF	DIAGNG	Read out display data stored on tape unit 3
REWIND	DIAGNG	Rewind tape unit NVAL
SAVEF	DIAGNG	Write display data on to tape unit 3
SPDIS	PROSCL	Display the display data array on the CRT
SPECT	PROSCL	Call SPECTO display routine
STAND	PROSCL	Produce a standard Spectrograph display
THRSP	DATGCL	Call THRSPIC data processing routine
WHATN	PROSCL & DATGCL	Call WHATNOW subroutine
ZEROC	PROSCL	Call ZEROC display routine.

Table 2 Commands Executed by Speech System

There are two basic formats for the output data. The three dimensional linear time displays are generally represented as a two-dimensional FORTRAN array (stored in FINT) with each element containing a quantity representing the intensity of the corresponding point on the display. The display routines can then normalize the data (performing such operations as high frequency emphasis, if desired), interpolate between data points and produce a smoothly varying, multi-intensity level display.

The x-y type of displays are represented as two arrays of the corresponding x and y coordinates of successive points in the display. These points can then be displayed as a continuous line using other system display routines. In addition other variants can be produced. In particular, a trivial modification of the above display program allows a single variable array to be plotted against time (i.e. successive values of the array are plotted vs. equidistant intervals on the x axis).

6.4 The Subprocessing Routines

The subprocessing routines consist of the programs which are used to perform various operations and transformations of data. Each routine performs a single type of operation and might be used in the construction of several different displays.

In order to insure their flexibility of use, the subprocessing routines have all been programmed to conform to a certain general form. In particular, each program receives as its input a data array and a variable indicating the number of points to be processed. The output may or may not be an array. If it is an array and if the output array contains the same number of points as the input, the program is written so that the same array can function as both input and output, if desired.

If the number of points in the output array is different from the number of input points, this number is specified as an output parameter.

In general, all intermediate data arrays used in the processing of data in the subprocessing routines are specified as parameters. This allows the calling programs to have complete control over the storage allocation of arrays and results in a considerable savings in space.

In order to avoid the variety of problems created by passing subroutine parameters through COMMON, this practice was generally not used. By passing all of the parameters explicitly, the routines are easier to understand and have many fewer mysterious side effects. There are two exceptions to this rule, however. One is that certain system constants were allowed to be obtained directly from COMMON, e.g. the sampling frequency, etc. In general, the variables which are passed in this manner are those whose use and meaning are unlikely to change as the system matures. This lowers the probability of having to rewrite the subprocessing routine later on. The second exception involves short subroutines which are used very often, i.e. in "tight loops". In such cases the overhead involved in handling explicit parameters becomes excessive so that passage through the COMMON area becomes necessary.

6.5 Basic System Principles

As the Speech Display Simulation System developed, certain key principles were developed as follows:

- 1) The common data base, command processor and speech display routines should be basically machine independent. This means that they should be written in standard FORTRAN as much as possible and any use of CSL FORTRAN extensions should be fully documented by means of comment statements in the code itself.

2) The subprocessing routines may be written in machine language or in a combination of FORTRAN and machine language as is allowed in the CSL FORTRAN system. However, this should only be done if a significant speedup in time or savings in space results or if it is necessary to perform some special function, such as communicating with the CRT display unit. In either case all occurrences of machine code should be explained both in the overall sense and at the detailed instruction level by comments within the program.

3) Test programs used to check out the various subprocessing routines are not normally to be loaded with the rest of the system. They are kept on the library tape, however, so that when needed, they may be easily loaded by making a call request to the CSL Operating System. These programs should be well commented with exact instructions on their use since it is easy to forget their operation within a matter of weeks if they are not used regularly.

The complete descriptions of the various programs used in the Speech Display system are given in Nordmann[1971] along with the program listings, test programs and sample outputs.

Chapter 7

RESULTS

The basic simulation system has worked quite well and proved quite adaptable as time went by. The major problem with the system at the present time is the amount of inconvenience involved in producing a digital data tape which can be used by the processing routines. Although the recording and playback through the A to D convertor is easy enough, the decision about what to save and put on the permanent data tape must be done on an individual basis by the operator. There are routines which can be used to assist in this operation, such as THRSPIIC which will print out, for each block on a tape, the number of samples above any particular threshold value chosen by the operator. However, the basic decision as to where the word starts and ends must be made by the operator. In a real time system it would be possible to get around such a problem by using a push button to indicate when you want the computer to "listen". At any rate, at the present time, this task is somewhat tedious. Once it is accomplished, however, the production of the displays is fairly simple.

The testing of the displays turned out to present quite a few problems, mostly revolving around the expense of a **really** comprehensive testing procedure. In the end it was necessary to restrict the amount of testing done with the result that the tests which were performed cannot in any way be considered definitive. However, several procedural variations were tested and certain generalizations can be made about the restricted tests which were performed.

In the end it was only possible to get two subjects who were able to complete a full set of tests and even these subjects were not able to run a full series on every display type. In addition several other

subjects completed various parts of the test series for specific display types. As a result it is impossible to make any statistically significant generalizations about the results and no type of statistical analysis was even attempted. It is hoped however that the results will prove useful in indicating the types of tests which might be useful in the future.

7.1 Recordings

The first area which became restricted was the recorded data itself. In order to minimize the number of utterances to be processed, the test vocabulary was restricted to the 40 words listed in Table 3.

The words were chosen so as to give a distribution over the full range of vowel sounds and at the same time allow maximum testing between words differing by only a single phoneme. Four speakers were used, three female and one male, to produce a total of 160 utterances. It was also intended to use a set of recordings of the Modified Rhyme Test (see Kreul, et. al [1968] or Beyer, et. al. [1969]) produced by the Stanford Research Institute and available from K-G Recording Service, 4311 Miranda Ave., Palo Alto, California. These recordings were originally produced to be used in speech discrimination tests but they were felt to be appropriate for the present purpose. Unfortunately a variety of equipment difficulties, some of which were never solved, prevented their conversion to digital tape. The result was that the number of utterances available for the second type of test was not really large enough.

The recordings of the 40 word list were produced in a quiet room using untrained friends of the author as speakers. The equipment used consisted of an Allied M3310 cardioid microphone attached to one channel of an Allied T-1070 stereo tape recorder. The use of untrained speakers produced one rather severe problem which was not discovered until several trial test runs had been performed, namely the words were not all enunciated clearly.

shin	nob	gin
beet	zed	dame
dead	bang	file
hag	cuff	stuck
sod	june	ned
four	thor	lip
mob	cage	wang
guff	lynn	hose
ted	pang	tame
sore	vile	rip
thin	wage	mile
shod	said	rang
peat	loon	gag
	chuck	

Table 3 List of Recorded Words

This caused confusions between certain particular utterances by certain speakers independent of the type of display used since the recordings themselves were ambiguous. The effect of this problem will be discussed further in the subsections concerning the actual test results.

7.2 Data From the First Test

As described in Section 4, the first test was intended to help determine if it was possible to extract the necessary information from a given type of display to identify different words consistently. It was also intended to give a measure of the relative efficiency with which the various display types performed this task by measuring the length of time needed to reach a certain proficiency with the display.

The test items for the first test were selected from the list of 40 words which were spoken by the 4 speakers. Two separate groups of items were used; the first (test 1a) consisting of the words zed, said, vile, file, dame, and tame and the second (test 1b) consisting of the words cuff, guff, mob, knob, shod, sod, ned and ted. The words in the two groups were chosen so as to provide pairs of words which might be easily confused if the displays were not in fact providing the proper cues. Unfortunately with the limited amount of testing which could be done, it was not possible to test for the full range of confusions between all the various phonemes.

The procedure for the first test involved showing the subject slides of the displays produced by a particular display type and having the subject try to determine which word was being displayed. When the subject responded, he was told whether or not his response was correct and if not, what word was actually being displayed. Initially the subject was allowed to look for five minutes at a labelled sheet containing pictures of all the slides in the test. Then the complete set was shown to him one at a time for

as many times as was necessary for it to be learned. During the test the subject was allowed to use a written list containing the words in the group being displayed.

Measurements were taken of the number of trial sets necessary to reach the criterion level of response. This level was loosely defined as the point at which the subject began to level off in improvement and started making a more or less consistent set of mistakes. It was more specifically specified as four consecutive trial sets in which the number of responses did not vary by more than 10%. Tables 4, 5, and 6 give the learning rates of each subject for the spectrographic, zero-crossing, and formant extracting displays, respectively, in terms of the number of trial sets necessary to reach the criterion run and the average percentage correct during the criterion run.

Confusion matrices were also constructed using the test results. By keeping the effects of the various speakers separate from one another, it was possible to determine effects which might be due to a single speaker alone. Tables 7 through 19 give the confusion matrices for each subject during their criterion runs, arranged in order of the type of display. Each box in each matrix has room for five numbers. The upper and lower left hand corners contain the number of times a particular response was given for display instances of the particular word as it was spoken by speakers a and b respectively. The upper and lower right hand corners contain the number of responses for instances involving speakers c and d respectively. The number in the center position is simply the sum of the numbers in the four corners and represents the total number of times a particular response was given to a display instance representing the particular word irrespective of which speaker pronounced it.

SUBJECT	SETS TO CRITERION	% CORRECT DURING CRITERION RUN	
A	4	92%	} Test 1a
B	1	72%	
C	1	65%	
D	0	62%	
E	4	60%	
A	3	84%	} Test 1b
B	3	90%	

Table 4 Learning Rates for
Spectrographic Display

SUBJECT	SETS TO CRITERION	%CORRECT DURING CRITERION RUN	
A	4	81%	} Test 1a
B	2	69%	
A	3	89%	} Test 1b
B	1	97%	

Table 5 Learning Rates for
Zero-Crossing Display

SUBJECT	SETS TO CRITERION	% CORRECT DURING CRITERION RUN	
A	4	92%	test 1a
A	3	93%	test 1b

Table 6 Learning Rates for
Formant Extracting Display

	ZED	SAID	VILE	FILE	DAME	TAME
ZED	4 4 13 4 1	3 3				
SAID	1 1	4 4 15 3 4				
VILE			4 3 15 4 4	1 1		
FILE				4 4 16 4 4		
DAME					4 4 15 4 3	1 1
TAME					2 2	4 2 14 4 4

Table 7 Confusion Matrix for Subject A, Test 1a, Spectrographic Display

	ZED	SAID	VILE	FILE	DAME	TAME
ZED	3 3 13 4 3	1 1 3 1				
SAID	2 3 6 1 4	2 1 10 3				
VILE			1 4 9 4	3 7 4		
FILE			3 4 3 13 4 4	1		
DAME					3 4 10 3	1 6 1 4
TAME					2 3 1	4 2 13 3 4

Table 8 Confusion Matrix for Subject B, Test 1a, Spectrographic Display

	ZED	SAID	VILE	FILE	DAME	TAME
ZED	4 1 12 4 3	3 4 1				
SAID	4 3 9 2	1 7 2 4				
VILE			1 5 4	3 7 4		
FILE			1 3 2 4	3 4 13 2		
DAME					4 4 14 4 2	2 2
TAME					3 4 4 1	4 1 12 3 4

Table 9 Confusion Matrix for Subject C, Test 1a, Spectrographic Display

	ZED	SAID	VILE	FILE	DAME	TAME
ZED	4 0 10	4 6				
	3 3	1 1				
SAID	2 1 4	2 3 12				
	1	3 4				
VILE			4 5			
			0 1	4 1	2	
FILE			3 1 6	1 3 10		
			1 1	3 3		
DAME				1	2 3	2
				5	6	5
				4	0 1	3
TAME					2	4 2
					3	13
					1	3 4

Table 10 Confusion Matrix for Subject D, Test 1a, Spectrographic Display

	ZED	SAID	VILE	FILE	DAME	TAME
ZED	2 0 6 2 2	2 4 9 2 1			1	
SAID	1 1 5 3	3 3 7 1 0	1 1	1 1	1	1
VILE	1	1	4 1 12 4 3	1 2 1	1	
FILE			3 4 6 3 4	1 4 10 1		
DAME					4 7 0 3	4
TAME						4 4 15 3 4

Table 11 Confusion Matrix for Subject E, Test 1a, Spectrographic Display

	CUFF	GUFF	MOB	NOB	SHOD	SOD	NED	TED
CUFF	5 15 5	3 4 2	2 2				1	
GUFF	1 5 1	3 13 4	5 1				1 2 1	
MOS			5 17 5	2 5			3 3	
NOB				5 20 5	5			
SHOD					5 20 5	5		
SOD						5 20 5	5	
NED		1	1 4 2	1			4 15 5	3
TED							2 2	5 18 3

Table 12 Confusion Matrix for Subject A, Test 1b, Spectrographic Display

	CUFF	GUFF	MOB	NOB	SHOD	SOD	NED	TED
CUFF	3 13 4	3 1 2 3	1 1 2					1 1
GUFF	1 1	3 15 4	4 4					
MOS			4 15 4	3 1				
NOB			1	3 10 3	4 X		1 1	
SHOD					4 16 4	4 4		
SOD						4 16 4	4 4	
NED			2 5 1	2			2 11 3	2 4
TED							1	4 15 4

Table 13 Confusion Matrix for Subject B, Test 1b, Spectrographic Display

	ZED	SAID	VILE	FILE	DAME	TAME
ZED	5 4 16 4 3	1 4 1 2				
SAID	1 1	4 5 19 5 5				
VILE			3 3 14 3 5	2 2 6 2		
FILE			2 6 2 2	3 5 14 3 3		
DAME					2 5 17 5 5	3 3
TAME					2 5 2 5	5 5 18 3

Table 14 Confusion Matrix for Subject A, Test 1a, Zero-crossing Display

	ZED	SAID	VILE	FILE	DAME	TAME
ZED	4 3 14 2 4 3 1	1 2 1 3 3				
SAID	3 4 1 9 7 1 1 3 3					
VILE			1 2 10 5 4 3 1	3 2		
FILE			1 1 6 10 2 2 2 2	3 3		
DAME					4 4 15 4 3 1	1
TAME					3 4 6 10 3 1 4	1

Table 15 Confusion Matrix for Subject B, Test 1a, Zero-crossing Display

	CUFF	GUFF	MOB	NOB	SHOD	SOD	NED	TED
CUFF	5 19 5	5 4						
GUFF		4 18 5	5 4				1	1
MOS			5 17 4	4 2 1			1	
NOB			3 3	5 11 5	1 X		1	
SHOD					5 20 5	5		
SOD						5 20 5		
NED							5 20 5	
TED					1		5 5 4	5 14 5

Table 16 Confusion Matrix for Subject A, Test 1b, Zero-crossing Display

	CUFF	GUFF	MOB	NOB	SHOD	SOD	NED	TED
CUFF	4 15 4 3	4 1 4 1						
GUFF	1 1	4 3 15 4						
MOB			4 15 4 3	1 1				
NOB				4 12 4 X				
SHOD					4 16 4 4			
SOD						4 16 4 4		
NED							4 16 4 4	
TED							1 1	4 3 15 4

Table 17 Confusion Matrix for Subject B, Test 1b, Zero-crossing Display

	ZED	SAID	VILE	FILE	DAME	TAME
ZED	5 20 5	5 1				
SAID	1	4 19 5	5			
VILE			5 19 5	4 1		
FILE			1	5 19 5	4	
DAME					5 20 5	5
TAME					5 6 1	5 14 4

Table 18 Confusion Matrix for Subject A, Test 1a, Formant Extracting Display

	CUFF	GUFF	MOB	NOB	SHOD	SOD	NED	TED
CUFF	5 5 18 5 3						1	
GUFF		5 5 20 5 5						
MOS			5 5 20 5 5					
NOB				3 5 3 13 5				
SHOD					5 5 20 5 5			
SOD						1 4 6 9 2 3 3 2		
NED							5 5 19 4 5	
TED								5 5 20 5 5

Table 19 Confusion Matrix for Subject A, Test 1b, Formant Extracting Display

As might be expected, the small number of subjects has caused a great deal of confounding of data since many of the possible sources of variance could not be balanced. In particular, the order in which a subject learned the displays and the order in which the parts of the first test were given could not be varied in such a way as to cancel any possible variance which might be due to learning effects (across displays as well as during the learning of a single display). As far as the learning data is concerned, there is a contradiction between the two parts of the first test which were performed on the different display types. The number of sets necessary to reach the criterion run and the percentage correct during the criterion run as recorded during test 1a would seem to indicate that the spectrographic display was easier to learn than the zero-crossing display. This same data on test 1b, however, tends to indicate the opposite. The most likely explanation appears to be that the differences in both cases are not large enough to be statistically significant given the small amount of data available.

The confusion matrix data shows several interesting points. In test 1a, there are very few confusions outside of the three basic word pairs, i.e. zed-said, vile-file, and dame-tame. This could be attributed to the fact that all three displays were able to satisfactorily distinguish vowels. More probably, however, it is due to the fact that the word pairs picked for this test have many differences among themselves and thus there are many cues available with which to distinguish them. A much more selective test could have been devised if the word pairs had been more similar in their phonemic structure, e.g. if they all ended in the same phoneme and used the same middle vowel.

A slight example of the type of results which this improvement might produce is shown in the confusion matrices for test 1b for the

spectrographic and zero-crossing displays. Both subject A and subject B had a certain amount of trouble distinguishing between the words "ned" and "mob" in the spectrographic display (refer to tables 12 and 13). However, there was no such problem with the zero-crossing display (see tables 16 and 17).

As can be seen from the various confusion matrices, there does appear to be a differential effect in the confusions of some of the test words based on which speaker's recordings were being used, e.g. "zed" in the spectrographic matrices for subjects C, D, and E is mistaken for "said" much more often in the case of speaker c than for any other speaker. It turned out that in the original recording it is in fact rather difficult to determine whether the word is a "zed" or a "said". The problem comes when we note that subjects A and B did not make this confusion.

This contradiction was eventually resolved by a comment one of the subjects made in regard to another similar situation, namely the word "vile" spoken by speaker b. This case was one in which there was a definite problem of consistent classification, but the particular display was so strikingly different that the subject simply put it in a class by itself and after missing it once, he never misclassified it again. This same effect probably occurred in other cases as well and if common, it would not only obscure the effects of poor pronunciation, but it would also tend to invalidate the tests, since the subjects would be memorizing specific instances instead of general identifying principles. The main way to correct this problem would be to have more speakers and have several different examples from each speaker. Then by having successive test sets composed of different instances, the subjects would never be able to memorize specific instances.

7.3 Data from the Second Test

The second test was intended to be a closer approximation to the final learning situation since it would involve a comparison between two displays shown simultaneously. Its purpose was to obtain more detailed data on the effectiveness of the displays and on the tolerances which were involved in each type.

Unfortunately, the Modified Rhyme Test recordings were found to be defective when played through the digital conversion apparatus of the display system. Thus it was necessary to use the same set of recordings as in the first test. But since there were not nearly enough instances in these recordings for a complete test, only a single test involving comparisons between seven words was attempted ("zed", "said", "ned", "ted", "sod", "shod", and "dead").

The actual procedures used in the test became rather complex due to some of the external restrictions placed on the experiment. Due to time and cost constraints, only one slide of each display instance was available. Thus an elaborate scheme had to be worked out whereby all possible comparisons of different instances could be performed with a minimum amount of slide shuffling between the two projectors. This was done by dividing the slides into two groups and working out all possible comparisons between the instances in the two groups. In order to keep the expectations of the two possible responses ("same" and "different") equal, it was necessary that approximately half of the matches in each set be the same word.

The various possible comparisons were written on small index cards and shuffled to give a semi-random ordering. Then the two sets of cards were placed in their respective projectors. One set was arranged so

that the experimenter could project the slides in any arbitrary order. The other set was shuffled and then displayed one at a time by the subject in sequence after the experimenter first noted down their order on a piece of paper. As the subject projected each of his slides in turn, the experimenter would pick the corresponding slide from his projector based on the current index card notations being used and project it next to the first slide. The subject would then respond "same" or "different", the experimenter would answer "right" or "wrong", and then both would go to the next slide pair. When the set was completed, the subject's slides would be shuffled, the experimenter would select a new set of instances to match and the process would repeat.

Since there were only single copies of each slide it was necessary to rearrange the two display sets periodically to match other combinations which could not be obtained using the previous set divisions. By having two complete sets of slides this could be eliminated. It would also be possible to have longer test runs between shufflings and to lower the total number of runs necessary.

Tables 20 through 23 give the data recorded from the second test for the spectrographic and zero-crossing displays, respectively. Only two runs were made using this test and the data is shown in two forms: a detailed matrix showing the results of the comparisons of the display instances of particular speakers and a summary matrix showing the proportion of "same" responses. In the detailed comparison matrix the speakers are listed along the sides of the matrix for each word in the test. For each instance pair tested a letter will appear at the respective intersection in the matrix: a "d" if the response was "different", an "s" if the response was "same". If no letter appears, the pair was not tested

ZED	.65						
SAID	.40	.82					
NED	.00	.00	.77				
TED	.33	.00	.40	.94			
SHOD	.00	.00	.17	.00	1.0		
SOD	.17	.33	.00	.00	.00	1.0	
	ZED	SAID	NED	TED	SHOD	SOD	

Table 21 Summary Comparison Matrix for Subject A,
Test 2, Spectrographic Display

[illegible]

Table 22 Detailed Comparison Matrix for Subject A,
Test 2, Zero-crossing Display

ZED	.78						
SAID	.63	.78					
NED	.00	.00	.83				
TED	.13	.00	.13	.87			
SHOD	.00	.00	.00	.00	.83		
SOD	.57	.38	.00	.00	.13	.96	
DEAD	.17	.00	.33	.50	.00	.00	.64
	ZED	SAID	NED	TED	SHOD	SOD	DEAD

Table 23 Summary Comparison Matrix for Subject A,
Test 2, Zero-crossing Display

and if more than one appears it was tested more than once. In the case of pairs of instances representing the same word, the proportion of "same" responses is given, since there were too many cases to write out a letter for each one.

The biggest problem with this test was the length of time it took to perform it. Subject A worked a total time of about 6 hours on the test for the spectrographic display and was still able to only see approximately half of the total number of possible comparisons. A great deal of the time was taken up in the procedural problems mentioned above and a double set of slides would probably cut down the amount of time needed by a significant factor. However, the fact remains that the procedure still will take a great amount of time because of the large number of instance pairs which must be tested. The results from the comparison test show several interesting features. In general, the mistakes appear to be made on the same word pairs in both the zero-crossing and spectrographic display types although the spectrographic display has a higher error rate in almost every case. This would tend to indicate that although the subjects have trouble on the same type of comparisons (at least as far as the words which were tested are concerned), the zero-crossing display tends to allow the subject to resolve the differences more accurately. (It should be noted in regard to the problem of learning effects that subject A performed the test for the zero-crossing display first).

The detailed data from the comparison test agrees with the results from the first test in certain respects. In the cases where the same word spoken by two different speakers was presented, the subject tended to make errors on the same instances as in the first test. This effect is most noticeable in the case of "zed" spoken by speaker C. When this instance

was compared to either speaker a or speaker d's "zed" the subject made a high error rate, but he made a perfect score when speaker b was compared to a or c.

Chapter 8

SUMMARY AND CONCLUSIONS

The discussion of the experimental results can be broken down into two main areas: a discussion of the tests themselves and a discussion of the general ideas behind the testing.

8.1 Comments on the Tests Which Were Performed

Although the tests which were performed could not be used to establish reliable comparisons between the various display types due to the small number of subjects which were used, they did indicate several points about the procedures to be used.

In picking out the words to be used in the test, an attempt was made to select a variety of words which would contain all of the common phonemes (at least in the English language). In order to minimize the total number of words, however, the selection was restricted in such a way that most of the words in the list differed from one another in several ways. It was originally felt that the effects of single phonemes could be determined from a multi-variant analysis of data from the complete set of words. Unfortunately, the amount of motivation and work necessary to adequately perform on a test with 40 or 50 different word displays to remember is much more than the average subject will ever have. When smaller subsets of the word list are used, it is not possible to control all of the variance.

Thus one basic change which should probably be made in the word lists is to use nonsense consonant-vowel-consonant syllables and to pick these syllables in such a way as to have subsets in which each "word" differs in only one phoneme. In order to keep the total number of words in the data set at a minimum, it will be necessary to restrict the number

of vowels used in these subsets. The vowels could then be tested using a subset in which only they vary.

The recordings themselves should be made by trained speakers. In the current tests this was not done and the result was that it was sometimes difficult to tell whether the subject confused two different examples of a particular word because the display was ineffective or because the audio recordings of the words were not differentiable.

In addition many more recordings of each word should be used. In the current tests, the subjects viewed all the instances during each trial set and from the comments which were made, it was apparent that they were memorizing specific instances. In order to avoid this, it would be necessary to have enough instances so that several test sets could be run without repeating any instances.

8.2 Comments on the General Method

Above and beyond any purely technical problems with the tests themselves there are some more general problems involving the whole idea behind the tests. As was mentioned in Section 4, there is the problem of whether or not the subject can use a display to correct his pronunciation even if he is able to detect that there is a difference between his pronunciation and the standard. This can only be answered when a real-time display system is developed and tests can be conducted on-line with the system.

There are other problems as well. For one thing, the word identification type of testing which was used in this experiment is not exactly the same type of situation as will be needed in the final use of the system. It might very well be that the speech deformations encountered in training the deaf or teaching the pronunciation of a

foreign language are qualitatively different from the differences between the pronunciation of different words in a single language. In such a case, the present type of testing may be inappropriate insofar as determining the suitability of the various display types. This question can be solved by using the appropriate types of recordings and seeing if the results of the tests change in any way.

One other objection to this technique is the difficulty of applying it to the specialized displays which are often used in speech correction, such as pitch indicators, nasality indicators, etc. In principle these types of indicators could probably be tested using the present techniques and the displays could probably be generated quite easily by the system. However, in the case of this type of display, a much simpler testing method could probably be devised.

8.3 Summary

The purpose of the present project was to develop a computer speech display simulation system capable of generating a wide variety of speech displays from a recorded speech input. Eventually it is hoped that this will lead to a system whereby a person can obtain visual feedback as a corrective measure for word pronunciation. The basic system would involve two displays, one representing the subject's pronunciation of a particular word and the other representing a correct pronunciation of the word. A computer would be used to process the incoming speech and produce a display containing features highly relevant to correct pronunciation. The subject's task would be to detect differences in the two displays and to change his pronunciation so as to make them more similar.

After conducting an extensive literature search to determine the types of schemes which had previously been used to display speech sounds,

a basic interactive display system was programmed using the CSL's CDC 1604 computer-graphics facility. The system has been designed to be open-ended and currently can produce photographs of a variety of display types. Unfortunately, the system as it stands now cannot operate in real time due to the slowness of the CDC 1604.

The simulation system was used to produce examples of several different types of displays. These displays were used in a series of preliminary tests designed to develop techniques for comparing the effectiveness of various types of displays. Several corrections and refinements to the testing methods are discussed.

REFERENCES

- Abramson, Normajo, "Visual Aids for the Speech Correction of the Deaf and Hard-of-Hearing", M.A. Thesis, Emerson College, 1952.
- Anderson, F., "An Experimental Pitch Indicator for Training Deaf Scholars", J. Acoust. Soc. Am., Vol. 32, No. 8, August 1960, pp. 1065-1074.
- Barton, G. W. Jr., and Barton, S. H., "Forms of Sounds as Shown on an Oscilloscope by Roulette Figures", Science, Vol. 142, 1963, pp. 1455-1456.
- Bennett, W. R., "The Correlatograph - A Machine for Continuous Display of Short-Term Correlation", Bell System Journal, Vol. 32, 1953, pp. 1173-1185.
- Beyer, M. R., Webster, J. C., and Dague, D. M., "Revalidation of the Clinical Test Version of the Modified Rhyme Words", J. Speech and Hearing Research, Vol. 12, 1969, pp. 374-378.
- Biddulph, R., "Short Term Auto-Correlation Analysis and Correlatograms of Spoken Digits", J. Acoust. Soc. Am., Vol. 26, No. 4, July 1954, pp. 539-541.
- Bitzer, D. L., and Slottow, H. G., "The Plasma Display Panel - A Digitally Addressable Display with Inherent Memory", Fall Joint Computer Conference Proceedings-1966, Vol. 29, Sparten, Washington, D. C., 1966, pp. 541-547.
- Bobrow, D. G., and Klatt, D. H., "A Limited Speech Recognition System", Fall Joint Computer Conference Proceedings-1968, Vol. 33, Sparten, Washington, D. C., 1968, pp. 305-318.
- Bower, G. H. and Trabasso, T. R., "Concept Identification", in: Studies in Mathematical Psychology, Chap. 2, R. C. Atkinson (ed.), Stanford University Press, Stanford, California, 1965, pp. 32-94.
- Bridges, C. C., "An Apparatus for the Visual Display of Speech Sounds", Am. J. of Psychology, Vol. 77, No. 2, June 1964, pp. 301-303.
- Brigham, E. O., and Morrow, R. E., "The Fast Fourier Transform", IEEE Spectrum, Vol. 4, No. 12, December 1967, pp. 63-70.
- Bruner, J. S., Goodnow, J. J. and Austin, G. A., A Study of Thinking, Science Editions, Inc., John Wiley and Sons, New York, 1962.
- Campanella, S. J., Coulter, D. C., and Speaker, D. M., "Formant Tracking Speech Band-width Compression System Improvements", Melpar Inc., Tech. Report AFAL-TR-65-5, AD-461 490, March 1965.
- Cavanagh, Anita, "A New Audio-Visual Aid for Speech", The Volta Review, Vol. 53, No. 1, January 1951, pp. 12-13, 40-41.

Chang, S., Pihl, G. E., Essigman, M. W., "Representations of Speech Sounds and Some of Their Statistical Properties", Proceedings of the IRE, Vol. 39, No. 2, February 1951a, pp. 147-153.

Chang, S. H., Pihl, G. E., and Wiren, J., "The Intervalgram as a Visual Representation of Speech Sounds", J. Acoust. Soc. Am., Vol. 23, No. 6, November 1951b, pp. 675-679.

Cochran, William T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., Maling, G. C., Jr. Nelson, D. E., Rader, C. M., and Welch, P. D., "What is the Fast Fourier Transform?", IEEE Trans. on Audio and Electroacoustics, Vol. AU-15, No. 2, June 1967, pp. 45-55.

Cohen, Martin L., "The ADL Sustained Phoneme Analyzer", Am. Annals of the Deaf, Vol. 113, 1968, pp. 247-252.

Conner, J. Edward, "Evaluation of the Voice Visualizer as an Aid in Teaching Voice Improvement", D. Ed. Thesis, Boston University, 1955.

Cooley, James W., and Tukey, John W., "An Algorithm for the Machine Calculation of Complex Fourier Series", Mathematics of Computation, Vol. 19, No. 90, April 1965, pp. 297-301.

Coyne, A. E., "The Coyne Voice Pitch Indicator", Teacher of the Deaf, Vol. 36, 1938a, pp. 3-4, 100-103.

Coyne, A. E. "More About the Voice Pitch Indicator", The Volta Review, Vol. 40, No. 10, October 1938b, pp. 549-552, 598-599.

Davis, K. H., Biddulph, R., and Balashek, S., "Automatic Recognition of Spoken Digits", J. Acoust. Soc. Am., Vol. 24, No. 6, November 1952, pp. 637-642.

Delattre, P. C., Liberman, A. M., and Cooper, F. S., "Acoustic Loci and Transitional cues for Consonants", J. Acoust. Soc. Am., Vol. 27, 1955, pp. 769-773.

Dolansky, L. O., "An Instantaneous Pitch Period Indicator", J. Acoust. Soc. Am., Vol. 27, No. 1, January 1955, pp. 67-72.

Dolansky, L., Ferullo, R. J., O'Donnell, M. C., and Phillips, N. D., "Teaching Intonation and Inflections to the Deaf", Northeastern University, Cooperative Res. Proj. No. S-281, 1965.

Dolansky, L., and Phillips, N. D., "Teaching Vocal Pitch Patterns Using Visual Feedback From the Instantaneous Pitch-Period Indicator for Self-monitoring", Northeastern University, VRA Proj. No. 1907-S, October 1966.

Dreyfus-Graf, J., "Sur les Spectres Transitoires d' elements Phonetiques", Helvetica Physica Acta, Vol. 19, 1946, pp. 404-408.

Dreyfus-Graf, J. Schweig, "The Sonograph: Elementary Principles", Arch. Angen. Wiss. Tech., (in French), Vol. 14, December 1948, pp. 353-362.

Dreyfus-Graf, J., "Le Steno-Sonographe Phonetique", Technische Mitteilungen PTT, Vol. 28, No. 3, 1950, pp. 89-95.

Dreyfus-Graf, J., "Sonograph and Sound Mechanics", J. Acoust. Soc. Am., Vol. 22, No. 6, November 1950, pp. 731-739.

Dudley, H., and Gruenz, O. O. Jr., "Visible Speech Translators with External Phosphors", J. Acoust. Soc. Am., Vol. 18, No. 1, July 1946, pp. 62-73.

Ewing, G. D., and Taylor, John F., "Computer Recognition of Speech Using Zero-Crossing Information", IEEE Trans. on Audio and Electroacoustics, Vol. AU-17, No. 1, March 1969, pp. 37-40.

Fabian, Fredrick E., "Evaluation of the Voice Visualizer as a Supplementary Aid in the Correction of Articulation Disorders", E. Ed. Thesis, Boston University, 1955.

Flowers, J. B., "The True Nature of Speech - With Application to a Voice-Operated Phonographic Alphabet Writing Machine", Trans. Am. Inst. of Elect. Engin., Vol. 35, Pt. 1, 1916, pp. 213-248.

Focht, L. R. and Piotrowski, C. F., "Voice Sound Recognition", Philco Corp., Tech. Report No. RAD-TR-66-507, AD-802 997, October 1966.

Foulkes, J. D., "Computer Identification of Vowel Types", J. Acoust. Soc. Am., Vol. 33, No. 1, January 1961, pp. 7-11.

Fry, D. B., and Denes, P., "The Solution of Some Fundamental Problems in Mechanical Speech Recognition", Language and Speech, Vol. 1, Pt. 1, January - March 1958, pp. 35-58.

Gentleman, W.M., and Sande, G., "Fast Fourier Transforms - For Fun and Profit", Fall Joint Computer Conference Proceedings - 1966, Vol. 29, Spartan, Washington, D. C., 1966, pp. 563-578.

Gruenz, O. O. Jr., and Schott, L. O., "Extraction and Portrayal of Pitch of Speech Sounds", J. Acoust. Soc. of Am., Vol. 21, No. 5, September 1949, pp. 487-495.

Halle, M., Hughes, G. W., and Radley, J. P. A., "Acoustic Properties of Stop Consonants", J. Acoust. Soc. Am., Vol. 29, No. 1, January 1957, pp. 107-116.

Harris, C. M., and Waite, W. M., "Display of Sound Spectrographs in Real Time", J. Acous. Soc. Am., Vol. 35, No. 5, May 1963, p. 729.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., and Cooper, F. S., "Effect of Third Formant Transitions in the Perception of the Stop and Nasal Consonants", J. Acoust. Soc. Am., Vol. 30, 1958, pp. 122-126.

Heygood, R. C. and Bourne, L. E., "Attribute and Rule Learning Aspects of Conceptual Behavior", Psychological Reviews, Vol. 72, 1965, pp. 175-196.

House, A. S., Goldstein, D. P., and Hughes, G. W., "Perception of Visual Transforms of Speech Stimuli: Learning Simple Syllables", Am. Annals of the Deaf, Vol. 113, 1968, pp. 215-221.

Huggins, W. H., "A Note on Autocorrelation Analysis of Speech Sounds", J. Acoust. Soc. Am., Vol. 26, No. 5, September 1954, pp. 790-792.

Hughes, G. W., "The Recognition of Speech by Machine", Research Laboratory of Electronics, Mass. Inst. of Tech., Tech. Report 395, AD-268 489, May 1961.

Hughes, G. W., and Hemdal, J. F., "Speech Analysis", Purdue Research Foundation, Lafayette, Indiana, AF Project 5628, Final Report, TR-EE65-9, AFRL-65-68, AD 624 555, July 1965.

Jakobson, R., Fant, C. G. M., and Halle, M., "Preliminaries to Speech Analysis", Acoust. Lab., Mass. Inst. of Tech., Tech. Report No. 13, 1952.

Jakobson, R., and Halle, M., Fundamentals of Language, Mouton and Co., Gravenhage, Netherlands, 1956.

Johnson, J. B., "A Cathode-Ray Tube for Viewing Continuous Patterns", J. of Applied Physics, Vol. 17, No. 11, November 1946, pp. 891-894.

Kersta, L. G., "Amplitude Cross-Section Representation with the Sound Spectrograph", J. Acoust. Soc. of Am., Vol. 20, No. 6, November 1948, pp. 796-801.

Kock, W. E., and Miller, R. L., "Dynamic Spectrograms of Speech", J. Acoust. Soc. Am., Vol. 24, No. 6, November 1952, pp. 783-784.

Koenig, W., Dunn, H. K., and Lacy, L. Y., "The Sound Spectrograph", J. Acoust. Soc. Am., Vol. 18, No. 1, July 1946, pp. 19-49.

Koenig, W., and Ruppel, A. E., "Quantitative Amplitude Representation in Sound Spectrograms", J. Acoust. Soc. Am., Vol. 20, No. 6, November 1948, pp. 785-795.

Kopp, G. A., and Green, H. C., "Basic Phonetic Principles of Visible Speech", J. Acoust. Soc. Am., Vol. 18, No. 1, July 1946, pp. 74-89.

Kopp, G. A., and Kopp, H. G., "Visible Speech for the Deaf", Speech and Hearing Clinic, Wayne State University, Final Report, Office of Vocational Rehabilitation, Dept. of HEW, 1963a.

Kopp, G. A., and Kopp, H. C., "An Investigation to Evaluate Usefulness of the Visible Speech Cathode Ray Tube Translator as a Supplement to the Oral Method of Teaching Speech to Deaf and Severely-deafened Children", Wayne State University, Final Report, Grant No. RD-526, Office of Vocational Rehabilitation, Dept HEW, 1963b.

Kreul, E. J., Nixon, J. C., Kryter, K. D., Bell, D. W., and Lang, J. S., "A Proposed Clinical Test of Speech Discrimination", J. Speech and Hearing Research, Vol. 11, No. 3, September 1968, pp. 536-552.

Ladefoged, P., and Broadbent, D. E., "Information Conveyed by Vowels", J. Acoust. Soc. Am., Vol. 29, No. 1, January 1957, pp. 98-104.

Lenneberg, E. H., "Biological Foundations of Language", John Wiley and Sons, Inc., New York, 1967.

Lerner, Robert M., Research Laboratory of Electronics, Mass. Inst. of Tech., Quarterly Progress Report, January 15, 1952, p. 55.

Lerner, Robert M., "A Method of Speech Compression", ScD. Thesis, Dept. E. E., Mass. Inst. of Tech., 1959.

Lieberman, A. M., "Some Results of Research on Speech Perception", J. Acoust. Soc. of Am., Vol. 29, No. 1, January 1957, pp. 117-123.

Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M., "Perception of the Speech Code", Psychological Review, Vol. 74, No. 6, November 1967a, pp. 431-461.

Lieberman, A. M., Cooper, F. S., Studdert-Kennedy, M., "Why Are Spectrograms Hard to Read?", Haskins Laboratories, Quarterly Progress Report, April-June 1967b, pp. 1.1-1.12.

Lieberman, A. M., Delattre, P. C., and Cooper, F. S., "Some Cues for the Distinction Between Voiced and Voiceless Stops in Initial Position", Language and Speech, Vol. 1, Pt. 3, July-September 1958, pp. 153-167.

Lieberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J., "The Role of Consonant Vowel Transitions in the Perception of the Stop and Nasal Consonants", Psychological Monographs, Vol. 68, No. 8, Whole No. 379, 1954.

Lieberman, A. M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F. S., "Minimal Rules for Synthesizing Speech", J. Acoust. Soc. Am., Vol. 31, No. 11, November 1959, pp. 1440-1499.

Licklider, J. C. R., "The Intelligibility of Amplitude-Dichotomized, Time-Quantized Speech Waves", J. Acoust. Soc. Am., Vol. 22, No. 6, November 1950, pp. 820-823.

Licklider, J. C. R., and Pollack, I., "Effects of Differentiation, Integrations and Infinite Peak Clipping Upon the Intelligibility of Speech", J. Acoust. Soc. Am., Vol. 20, No. 1, January 1948, pp. 42-51.

Majewski, W., and Hollien, H., "Formant Frequency Regions of Polish Vowels", J. Acoust. Soc. Am., Vol. 42, No. 5, November 1967, pp. 1031-1037.

Martony, Janos, "On the Correction of the Voice Pitch Level for Severely Hard-of-Hearing Subjects", Am. Annals of the Deaf, Vol. 113, 1968, pp. 195-202.

Mathes, R. C., Norwine, A. C., and Davis, K. H., "The Cathode-Ray Sound Spectroscope", J. Acoust. Soc. Am., Vol. 21, No. 5, September 1949, pp. 527-537.

Noll, A. M., "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection", J. Acoust. Soc. Am., Vol. 36, No. 2, February 1964, pp. 296-302.

Noll, A. M., "Cepstrum Pitch Determination", J. Acoust. Soc. Am., Vol. 41, No. 2, February 1967, pp. 293-309.

Nordmann, Bernard J. Jr., "Speech Display Simulation System for a Comparative Study of Some Visual Speech Displays", Digital Computer Laboratory, University of Illinois, Tech. Report 470, August 1971; also as Coordinated Science Laboratory, University of Illinois, Tech. Report R-524, September 1971.

Oppenheim, Alan V., "Speech Spectrograms Using the Fast Fourier Transform", IEEE Spectrum, Vol. 7, No. 8, August 1970, pp. 57-62.

Otten, K. W., "Simulation and Evaluation of Phonetic Speech Recognition Techniques - Acoustical Characteristics of Speech Sounds Systematically Arranged in the Form of Tables", National Cash Register Co., RTD-TDR-63-4005, Vol. III, AD-601 422, March 1964a.

Otten, K. W., "Simulation and Evaluation of Phonetic Speech Recognition Techniques - Indexed Bibliography on Speech Analysis, Synthesis, and Processing", National Cash Register Co., RTD-TDR-63-4005, Vol. IV, AD-601 421, April 1964b.

Otten, K. W., "Simulation and Evaluation of Phonetic Speech Recognition Techniques - Summary Report", National Cash Register Co., RTD-TDR-63-4005, Vol. V, AD-602 691, April 1964c.

Peterson, G. E., "Design of Visible Speech Devices", J. Acoust. Soc. Am., Vol. 26, No. 3, May 1954, pp. 406-413.

Phillips, Nathan D., Remillard, Wilfred, Bass, Susan, and Pronovost, Wilbert, "Teaching of Intonation to the Deaf by Visual Pattern Matching", Am. Annals of the Deaf, Vol. 113, 1968, pp. 239-246.

Pickett, J. M., ed., "Proceedings of the Conference on Speech-Analyzing Aids for the Deaf", Am. Annals of the Deaf, Vol. 113, 1968, pp. 117-326.

Pickett, James M., "Some Applications of Speech Analysis to Communication Aid for the Deaf", IEEE Trans. on Audio and Electroacoustics, Vol. AU-17, No. 4, Dec. 1969, pp. 283-289.

Pickett, J. M., and Constam, Alfred, "A Visual Speech Trainer with Simplified Indication of Vowel Spectrum", Am. Annals of the Deaf, Vol. 113, pp. 253-258.

Plant, G.R.G., "The Plant-Mandy Voice Trainer - Some Notes by the Designer", Teacher of the Deaf, Vol. 58, 1960, pp. 12-15.

Plomp, R., Pols, C. W., Van de Geer, J. P., "Dimensional Analysis of Vowel Spectra", J. Acoust. Soc. Am., Vol. 41, No. 3, 1967, pp. 707-712.

Potter, R. K., "Visible Patterns of Sound", Science, Vol. 102, No. 2654, November 9, 1945, pp. 463-470.

Potter, R. K., "Introduction to Technical Discussions of Sound Portrayal", J. Acoust. Soc. Am., Vol. 18, No. 1, July 1946, pp. 1-3.

Potter, R. K., Kopp, G. A., and Green, H. C., Visible Speech, D. Van Nostrand Co. Inc., New York, 1947.

Presti, A. J., "High Speed Sound Spectrograph", J. Acoust. Soc. Am., Vol. 40, No. 3, September 1966, pp. 628-634.

Prestigiacomio, A. J., "Plastic Tape Sound Spectrograph", J. of Speech and Hearing Disorders, Vol. 22, No. 3, September 1957, pp. 321-327.

Prestigiacomio, A. J., "Amplitude Contour Display of Sound Spectrograms", J. Acoust. Soc. Am., Vol. 34, No. 11, November 1962, pp. 1684-1688.

Pronovost, Wilbert, "Visual Aids to Speech Improvement", J. of Speech Disorders, Vol. 12, No. 4, December 1947, pp. 387-391.

Pronovost, W., "A Pilot Study of the Voice Visualizer for Teaching Speech to the Deaf", Proceedings of the International Congress on Education of the Deaf 1963, U. S. Government Printing Office, Senate Document No. 196, 1964.

Pronovost, Wilbert, Yenkin, Linda, Anderson, D. C., and Lerner, R., "The Voice Visualizer", Am. Annals of the Deaf, Vol. 113, 1968, pp. 230-238.

Pyron, B. O., and Williamson, F. R., Jr., "Visual Display of Speech by Means of Oscillographic Roulette Figures", Science, Vol. 145, 1964, pp. 72-73.

Pyron, B. O., and Williamson, F. R., Jr., "Study and Analysis of Signal Display and Bandwidth Compression Techniques", Georgia Institute of Tech., Final Report, Project A-791, Contract DA 49-092-ARO-52, AD 616 644, June 1965.

Radley, J. P., "The Role of Formant Transitions in the Identification of English Stops", MS Thesis, Mass. Inst. of Tech., 1956.

Ramaswamy, T. K., and Ramakrishna, B. S., "Simple Laboratory Setup for Obtaining Sound Spectrograms", J. Acoust. Soc. Am., Vol. 34, No. 4, April 1962, pp. 515-517.

Reddy, D. R., "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave", Ph.D. Thesis, Computer Science Dept., Stanford University, Tech. Report CS49, AD-640 836, September 1966.

Restle, F., "The Selection of Strategies in Cue Learning", Psychological Reviews, Vol. 69, No. 4, July 1962, pp. 329-343.

Riesz, R. R., and Schott, L., "Visible Speech Cathode-Ray Translator", J. Acoust. Soc. Am., Vol. 18, No. 1, July 1946, pp. 50-61.

Risberg, Arne, "Visual Aids for Speech Correction", Am. Annals of the Deaf, Vol. 113, 1968, pp. 178-194.

Risberg, A., "A Critical Review of Work on Speech Analyzing Hearing Aids", IEEE Trans. on Audio and Electroacoustics, Vol. AU-17, No. 4, December 1969.

Sakai, T. and Doshita, S., "The Automatic Speech Recognition System for Conversational Sound", IEEE Trans. on Electronic Computers, Vol. EC-12, No. 6, December 1963, pp. 835-846.

Sakai, T., Doshita, S., Niimi, Y., and Tabata, K., "Fundamental Studies of Speech Analysis and Synthesis", Am. Annals of the Deaf, Vol. 113, 1968, pp. 156-167.

Sakai, T., and Inoue, S., "New Instruments and Methods for Speech Analysis", J. Acoust. Soc. Am., Vol. 32, No. 4, April 1960, pp. 441-450.

Stark, Rachel E., Cullen, John K., and Chase, Richard A., "Preliminary Work with the New Bell Telephone Visible Speech Translator", Am. Annals of the Deaf, Vol. 113, 1968, pp. 205-214.

Steinberg, J. C., and French, N. R., "The Portrayal of Visible Speech", J. Acoust. Soc. Am., Vol. 18, No. 1, July 1946, pp. 4-18.

Stevens, K. N., "Autocorrelation Analysis of Speech Sounds", J. Acoust. Soc. of Am., Vol. 22, No. 6, November 1950, pp. 769-771.

Teacher, C. F., Kellett, H. G., and Focht, L. R., "Experimental Limited Vocabulary, Speech Recognizer", IEEE Trans. on Audio and Electroacoustics, Vol. AU-15, No. 3, September 1967, pp. 127-130.

Teacher, C. F., and Piotrowski, C. F., "Voice Sound Recognition", Philco, Corporation, Tech. Report RAD-TR-65-184, AD-619 964, July 1965.

Thomas, I. B., "The Significance of the Second Formant in Speech Intelligibility", Biological Computer Laboratory, Dept. of E. E., Univ. of Illinois, Tech. Report No. 10, July 1966.

Timberlake, Josephine B., "The Coyne Voice Pitch Indicator", The Volta Review, Vol. 40, No. 8, August 1938, pp. 437-439, 468-469.

Upton, Hubert W., "Wearable Eyeglass Speechreading Aid", Am. Annals of the Deaf, Vol. 113, 1968, pp. 222-229.

Vilbig, F., "An Apparatus for Speech Compression and Expansion and for Replaying Visible Speech Records", J. Acoust. Soc. Am., Vol. 22, No. 6, November 1950, pp. 754-761.

Vilbig, F., "Visible Speech-Rotary Field Coordinate-Conversion Analyser", IRE Trans. Audio, Vol. AU-2, No. 2, March-April 1954, pp. 76-80.

Williams, D. E., "A Visual Display of Certain Speech Parameters", MS Thesis, U. S. Naval Postgraduate School, AD-820 518, July 1967.

Willson, R. H., "A Capacitively Coupled Bistable Gas Discharge Cell for Computer Controlled Displays", Coordinated Science Laboratory, Univ. of Illinois, CSL Report R-303, June 1966.

Wood, D. E., and Hewitt, T. L., "New Instrumentation for Making Spectrographic Pictures of Speech", J. Acoust. Soc. Am., Vol. 35, No. 8, August 1963, pp. 1274-1278.

Wood, D. E., "New Display Format and a Flexible-Time Integrator for Spectral-Analysis Instrumentation", J. Acoust. Soc. Am., Vol. 36, No. 4, April 1964, pp. 639-643.

VITA

Bernard Joseph Nordmann Jr. was born in the little town of Lawton, Oklahoma on October 28, 1943, the eldest of five children born to Rosita and Bernard Nordmann Sr. After a boyhood spent in wandering the nations of the earth, he spent five strenuous years in the Boston metropolitan area satisfying the requirements for the S.B. and S.M. degrees (E.E.) from the Massachusetts Institute of Technology which he received in 1966. Searching for a change of scenery, he next journeyed to the fabled midwest where he settled in the mystical land of Champaign, Illinois. Here he worked for the Department of Computer Science on the Illiac III computer project while working towards his Ph.D. degree which he completed in 1971.

While working on the Illiac III project, Mr. Nordmann was responsible for the design of the main central processors used in the system. He also worked for the U.S. Naval Ordnance Laboratory in fits and starts between the years 1963 and 1966.

Mr. Nordmann is a member of the Institute of Electrical and Electronic Engineers, The Association for Computing Machinery and Sigma Xi.