SOME CONSIDERATIONS OF DEAF SPEECH

by

M. HELENA VERGARA NOLAN

B. S., Universidad Javeriana, Bogota, Colombia, 1971

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1974

Approved by

Nasir Ahmed
Major Professor

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The child who suffers from early deafness is faced with a doubly severe communication handicap. First, normal speech is unintelligible to him. Consequently there is a lack of exposure to speech in his early development and an inability to auditorily monitor his own vocalizations. These factors lead to the second handicap, namely, the great difficulty in learning to speak. The overall effect of these handicaps is that there is severe retardation of intellectual development. If better means can be found for overcoming the deficient speech communication, large improvements could occur in their education. The number of such persons in the United States is on the order of 400,000.

From the results of specialized teaching efforts, it follows that the second of the above handicaps can, in principle, be overcome. It is hoped that a much larger proportion of the deaf population could be trained to improve their ability to speak if appropriate instrumentation could be developed for helping the deaf in learning to speak. One approach to this problem has been the development of speech analyzing aids for speech training. Such aids typically operate on the acoustical signal, extracting that information which is not normally available to the deaf user. Such information is then presented to the deaf person either visually, tactually, or using whatever residual hearing is available.

In order to develop better speech analyzing aids for the deaf, it is necessary to find out more about the acoustic and perceptual characteristics of the speech of deaf children and how these characteristics differ from those of normal children. A survey of the literature shows that almost all of the efforts related to the study of acoustic characteristics have relief heavily on analog filter systems (e.g., a sound spectrograph), which are relatively imprecise compared to digital methods of analysis.

There are several advantages in using digital signal processing techniques. Analog techniques offer little flexibility for changing filter frequencies or the shape and width of the spectral windows used. Since there is no "optimum" compromise between the conflicting requirements of filter bandwidth and integration time, the need for choosing spectral windows is important, particularly in the case of deaf speech. A much more important advantage of digital signal processing techniques is that they provide the investigator the freedom in choosing the most appropriate method of analysis for each problem. Again, computerized techniques possess the capacity of handling large quantities of data. One of the problems encountered in analyzing the speech of the deaf is the large variability between speakers.

The principal objective of this report is to initiate an interdisciplinary effort between the Departments of Electrical Engineering and Speech of Kansas State University, in the area of speech processing of deaf speakers. Chapter II introduces the fundamental concepts of speech production, while differences between normal and deaf speech are presented in Chapter III. Some aspects of the digital signal processing of deaf speech are discussed in Chapter IV. Finally, some conclusions and recommendations for future work are presented in Chapter V.

CHAPTER II

SPEECH PRODUCTION FUNDAMENTALS

2.1 Introductory Remarks

The main objective of this chapter is to present a summary of some of the basic aspects of speech production. This material will be used to good advantage in the chapters that follow.

2.2 The Vocal Tract [1, 17]

The production of sounds in speech takes place in the vocal tract. The vocal tract is the region consisting of the larynx, pharynx, nose, mouth and lips. In the production of sounds, the shape of the vocal tract is changed by movements of the articulators which consist of the jaw, tongue, teeth, lips and velum. The vocal tract and the articulators are illustrated in Fig. 2.1-1 [17].

Acoustically, the behavior of the vocal tract is similar to that of a tube filled with air. If this air is disturbed, it will vibrate before returning to its rest condition. The natural frequencies of vibration of this air are called the formant frequencies. These formants have frequencies of F_1 = 500Hz, F_2 = 1500 Hz, F_3 = 2500 Hz and so on in steps of 1000 Hz. The values of these frequencies can change as a result of movements of the articulators. Fig. 2.2-2 shows the standing wave distribution pattern for

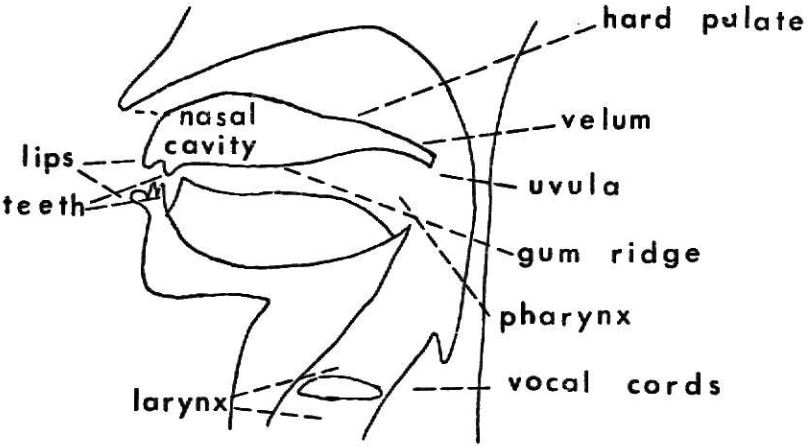

Figure 2.2-1 Human vocal apparatus

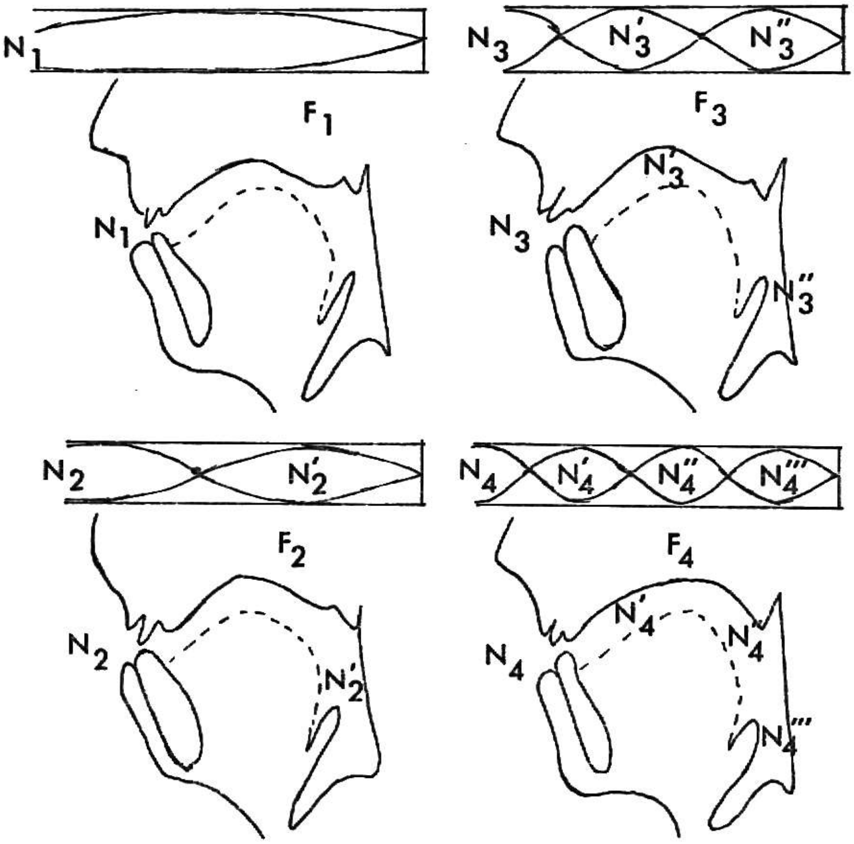

Figure 2.2-2 Standing wave distribution pattern

each of the first four resonances within a single tube representation of the vocal tract [11].

The sounds produced in speech have different ways of being generated. As a result they have variety within their characteristics. There are at least three different methods for these sounds to be produced:

a) If air molecules vibrate in a random way, the sound produced has a characteristic hissing quality.

b) When a flow of air is interrupted at regular intervals, a buzzing sound is produced.

c) If air under pressure is suddenly released, a plosive quality sound is achieved.

With different combinations of the above sounds and allowing variations on shape and size of the vocal tract, it is possible to generate phonemes, which are the basic units in speech production.

2.3 Acoustic Theory in the Production of Phonemes [1, 2]

One classification of phonemes is based on the way their sound source is produced. The arrangement of the phonemes in English language is as follows.

a) Vowels and Vowel-like consonants. Vowel-like consonants are shorter versions of the vowels. They are produced by vocal cord vibrations and a raised soft palate. In addition, resonances of the pharyngeal and oral cavities are involved.

b) Fricatives. Their sound source is air turbulence at a constriction. The spectral characteristic of the generated sound is governed by the size, shape, and length of the constriction.

c) Stops. Voiceless stops are produced by blocking the air flow in the vocal tract and allowing pressure to build up.

d) Affricates. Are produced by a stop phoneme followed by a turbulence which is deliberately extended by retaining a constriction.

e) Nasals. Three kinds of resonances are involved. These occur in the nasal cavity, the pharyngeal and in the oral cavity. The resonances in the nasal cavity are produced by lowering the soft palate.

In the sections that follow, the above classes of phonemes will be discussed in some detail.

2.4 Vowels [1, 3]

The source of excitation for the vocal tract during the production of vowels is a quasiperiodic series of pulses of air that pass through the glottis. Fig. 2.4-1 illustrates the system involved in the production of vowels, while Fig. 2.4-2 shows an equivalent lumped circuit. This circuit is valid at low frequencies and it is considered to represent the vocal tract during the production of vowels. This representation is valid only if the dimensions of the vocal tract are small compared with a wavelength in the frequency range of interest. In Fig. 2.4-2, C and M represent the volumes of the tract and the narrow mouth opening respectively.

2.5 Fricative Consonants [1, 4]

A narrow constriction characterizes the articulatory configuration. The position of the constriction depends upon the particular consonant. The procedure can be described as

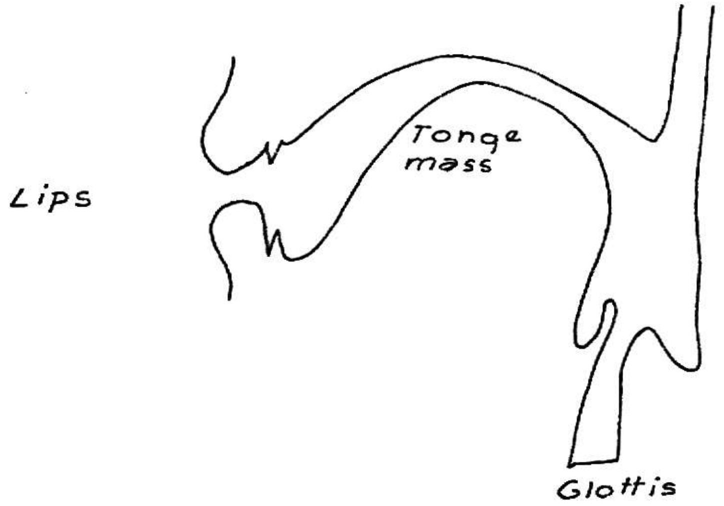

Figure 2.4-1 System involved in the production of vowels

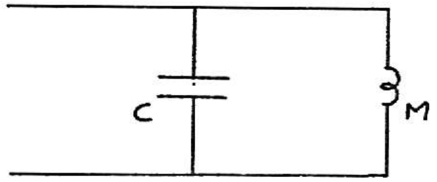

Figure 2.4-2 Equivalent lumped circuit for the production of vowels

follows: air is forced through this constriction at high velocity, and turbulent flow occurs in the vicinity of the constriction and possibly also at the teeth. Noise is generated as a result of the turbulent flow. This noise acts as exitation for the acoustic tube that forms the constriction and also for the cavities interior to the constriction. In addition, there can be some acoustical coupling through the constriction to the rear cavities. Fig. 2.5-1 shows an equivalent-circuit representation for the production of fricative consonants.

Some of the differences between this equivalent-circuit and the one for the vowels shown in Fig. 2.4-2 are as follows:

a) Here the source is a constant pressure noise source while that for the vowels is a periodic constant velocity source.

b) In this case the source is located within the vocal tract. In the case of the vowels, it is located at the glottis.

c) Additional damping is introduced by turbulence losses at the constriction. Also there are losses at the glottis since it remains somewhat more open than during vowel production, in order that air may flow continuously through the tract.

2.6 Nasals [1, 5]

Principally two essential features characterize the production of nasal sounds:

(a) The participation of the nasal passages in the formation of the spectral characteristics of the sound output.

(b) The oral cavity becomes a side branch of the main resonant tube [24].

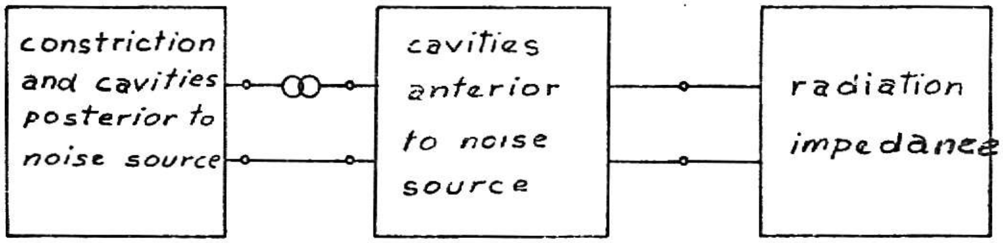
constriction and cavities posterior to noise source
cavities anterior to noise source
radiation impedance

Figure 2.5-1 Equivalent circuit representation for the production of fricative consonants

Figure 2.6-1 shows the entire articulatory system for the production of nasal sounds. This articulatory system is formed by three principal subsystems, namely, the pharynx, the oral cavity, and the nasal tract.

The spectra of nasal sounds may vary considerably from one sample to another. It depends upon the individual nasal consonant and its context. These spectra have the following three features:

(a) Very low first formant which is located at about 300 Hz and it is well separated from the upper formant structure. This feature results in a relatively well defined low concentration of energy.

(b) High damping factors of the formants.

(c) High density of the formants in the frequency domain.

As a consequence of the last two of the above features, the sound energy in the middle-frequency range (800-2300 Hz) is almost constant.

2.7 Speech Sounds in Context [1].

There are a variety of features to be considered in the study of the speech sounds in context. However, only a few of these are considered in what follows.

When sound is produced in connected speech, it may show the influence of the preceding or following sound. This occurs because the articulators cannot be in two places simultaneously. They take a finite time to move from the

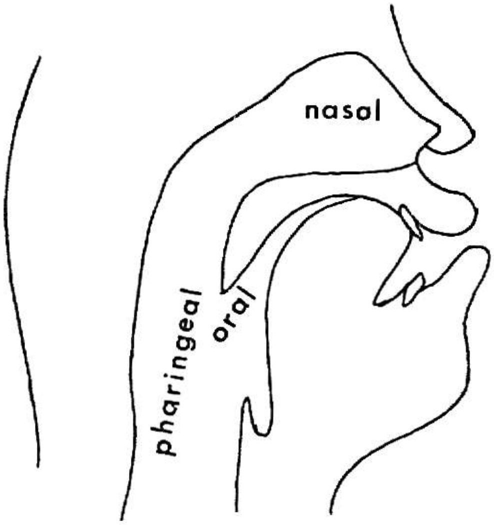

Figure 2.6-1 Articulatory system for the production of nasal sounds

configuration of one sound to the configuration of the next. Among the principal characteristics of the production of speech sounds in context are the transitions. These are formant changes and have several causes. For example, formant F_1 transitions result from change of tongue height. Tongue position changes are reflected in formant F_2. Again, formant F_3 transitions occurs when the size of the lip opening changes.

It is important to emphasize that the speech process does not involve a sequence of fixed articulatory configurations. It involves a sequence of movements, each movement conveying information about where the articulators have been and where they are headed.

Individual phonemes are related to the segmental structure of the speech. On the other hand, entire phrases are related to the suprasegmental structure of speech. The suprasegmental structure affects the meaning of what is said and also the voice quality. Features of suprasegmental structure that affect meaning are the intonation, stress, rhythm and phrasing.

Intonation is the modulation of voice pitch, where pitch is a subjective auditory percept. The intonation is determined by the frequency of vibration of the vocal cords. It is also affected by the air pressure developed by the lungs. The basic unit for characterizing intonation contours is the breath group. The breath group is the position of an utterance between pauses for taking breath.

To stress a syllable is to speak in such a way as to make it perceptually more prominent.

The pattern in which certain syllables are stressed and others are left unstressed in called rhythm.

Phrasing consists of grouping words together according to the linguistic structure of the utterance. A basic unit in phrasing is the sense group which is a group of one or more words forming a distinct linguistic entity. An example of a sense group could be a phrase or clause. Neighboring phonemes affect the production of individual phonemes. Phonemes within a sense group interact. There is no interaction between adjacent phonemes belonging to separate sense groups.

CHAPTER III

DIFFERENCES BETWEEN NORMAL AND DEAF SPEECH

3.1 Introductory Remarks

As it is known, the function of the hearing system in speech is to autocontrol the generation of the correct acoustical patterns. Consequently, differences in the speech of a deaf person are inherent in the absence of acoustic feedback.

3.2 Differences with Respect to Spectrographic Analysis

Before studying differences between normal and deaf speech that have been derived via a spectrographic analysis, it is instructive to consider some fundamentals related to this type of analysis.

Spectrographic analysis is a process by which sounds are analyzed to show how much energy they contain at various frequencies. These measurements recorded in visible form provide a spectrogram, some aspects of which are discussed next.

For the purposes of discussion, consider a simplified block diagram for the production of vowels, as shown in Fig. 3.2-1 [8]. The dc-air flow from the lungs is converted into ac-acoustic energy by the vocal cords whose spectrum is modified by the resonances of the vocal tract. The output

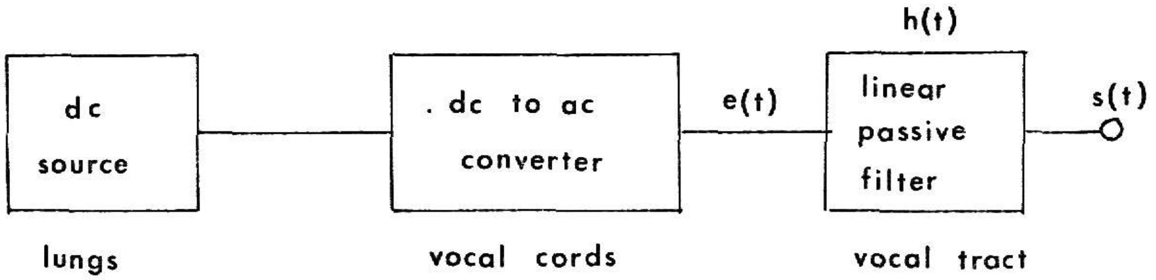

Figure 3.2-1 Block diagram representing the production of vowels

speech signal s(t) appearing at the lips is the convolution of the excitation function e(t), representing the air flow at the vocal cords, with the impulse response of the filter representing the vocal tract:

$$s(t) = \int_{-\infty}^{t} e(\tau) \cdot h(t - \tau) d\tau \qquad (3.2\text{-}1)$$

Taking the Fourier transform (designated by capital letters) of both sides of Eq. (3.2-1), one obtains

$$S(f) = E(f) \cdot H(f) \qquad (3.2\text{-}2)$$

Thus, the spectrum of the speech signal is the product of the excitation spectrum E(f) and the transfer function of the vocal tract H(f).

By taking absolute values, one obtains the amplitude spectrum of the speech signal,

$$|S(f)| = |E(f)| \cdot |H(f)| \qquad (3.2\text{-}3)$$

In actual speech, both the excitation spectrum and the transfer function vary with time. Thus, infinite time spectral analysis has to be replaced by short-time spectral analysis [12], [13], employing a time window comparable in duration to the shortest speech sounds. In practice, to preserve adequate spectral resolution, time windows with a width of up to 40 ms are often used, although some speech sounds may be somewhat shorter.

The results of short-time analysis depend on running time which implies that the spectra become functions of time. This is accounted for in the notation

$$|S(f,t)| = |E(f,t)| \cdot |H_t(f)|, \qquad (3.2\text{-}4)$$

where $|H_t(f)|$ is the transfer function of the vocal tract at time t. If the transfer function can be considered stationary during the analysis period, Eq. (3.2-4) is an exact relation, otherwise it is only an approximation. The goodness of the approximation depends on the stationarity of $|H_t(f)|$. For normal speaking rates and analysis periods not exceeding 40 ms, the effects of time varying transfer functions can usually be neglected.

From Eq. (3.2-4) it follows that $20 \log_{10} |S(f,t)| = 20 \log_{10} |E(f,t)| + 20 \log_{10} |H_t(f)|$. In this form, the effects of the excitation function and the transfer function are additive.

For example, Fig. 3.2-2 shows the short-time spectrum of the utterance "I triple E Proceedings" uttered by an adult male speaker of American English [8]. A time-window of about 3 ms duration was used. Time runs along the abscissa, frequency along the ordinate. Different shades of grey signify different spectral intensities, black corresponding to the highest value of $|S(f,t)|$ and white to the lowest. The curving black bands represent the formant frequencies. The irregularly striped patches indicate aperiodic (noise-like) exitation for unvoiced speech sounds.

Figure 3.2-3 shows a spectrogram of the same utterance but with a longer analysis time (20 ms) and correspondingly higher spectral resolution ($\Delta f = 50$ Hz). As a result, the

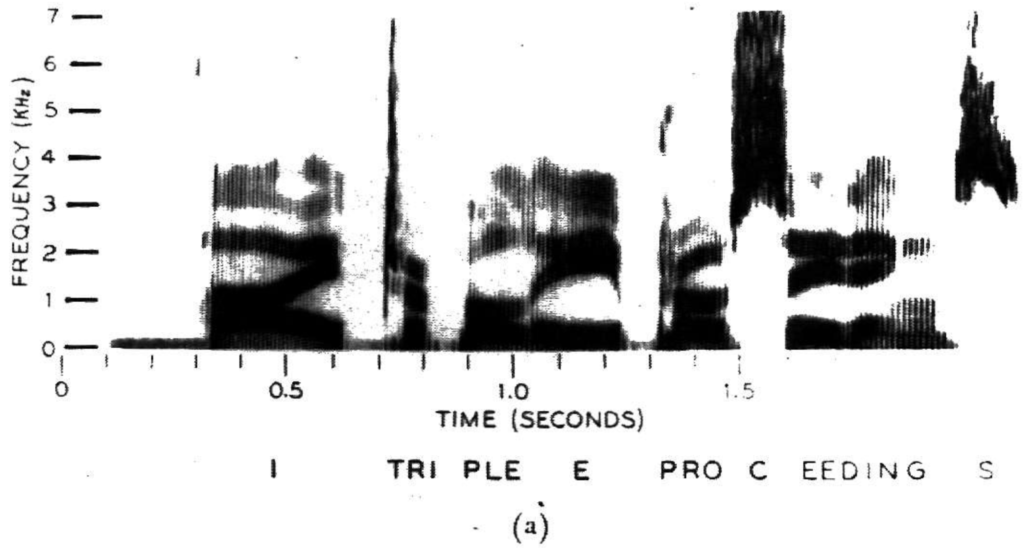

Figure 3.2-2 Speech spectrogram of the utterance "I triple E PROCEEDINGS" (wide-band analysis)

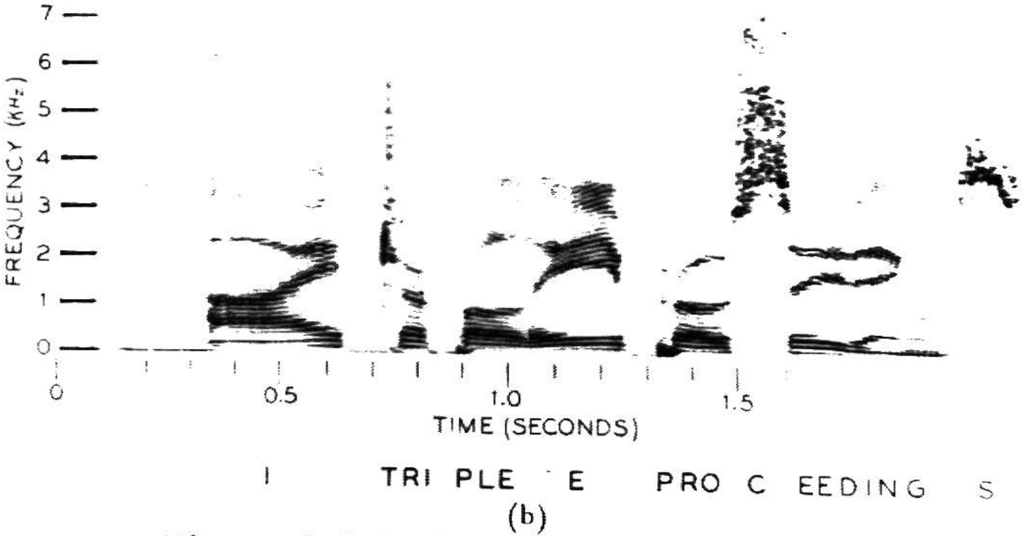

Figure 3.2-3 Speech spectrogram of the utterance "I triple E PROCEEDINGS" (narrow-band analysis)

individual harmonics of voiced sounds become visible in the spectrogram (the nearly horizontal narrow lines).

For a more detailed analysis of speech signals, the short-time spectrum at one instant in time or "section" may be obtained. Figure 3.2-4 is a section for the "ee" sound in "Proceedings". In Fig. 3.2-4 the abscissa is frequency (from right to left) and the ordinate is the logarithmic spectrum $10 \log_{10} |S(f,t_o)|$. The individual harmonics are visible as sharp spikes while the formants are represented by the smooth out line or "envelope."

Figure 3.2-4 suggests an alternative representation of the short-time spectrum, namely in terms of its spectral envelope $|G(f,t)|$ indicated by a dashed line, and the spectral fine-structure $|F(f,t)|$:$20 \log_{10} |S(F,t)| = 20 \log_{10} |F(f,t)| +20 \log_{10} |G(f,t)|$. For voiced speech sounds, $|F(f,t)|$ has equidistant maxima at the fundamental frequency f_o and its overtones or harmonics; for unvoiced sounds, $|F(f,t)|$ is a continuous function of frequency.

The following results have been derived by several investigators upon the spectrographic analysis. Among them there are some typical abnormalities, for example, the vowels are strikingly prolonged [14] and hence distorted. The formants are broad, not concentrated bands of energy, and considerable energy is spread throughout the full bandwidth of the spectrogram as in the case with many noises. The analyses reflect that the fundamental frequency and the formants do not change

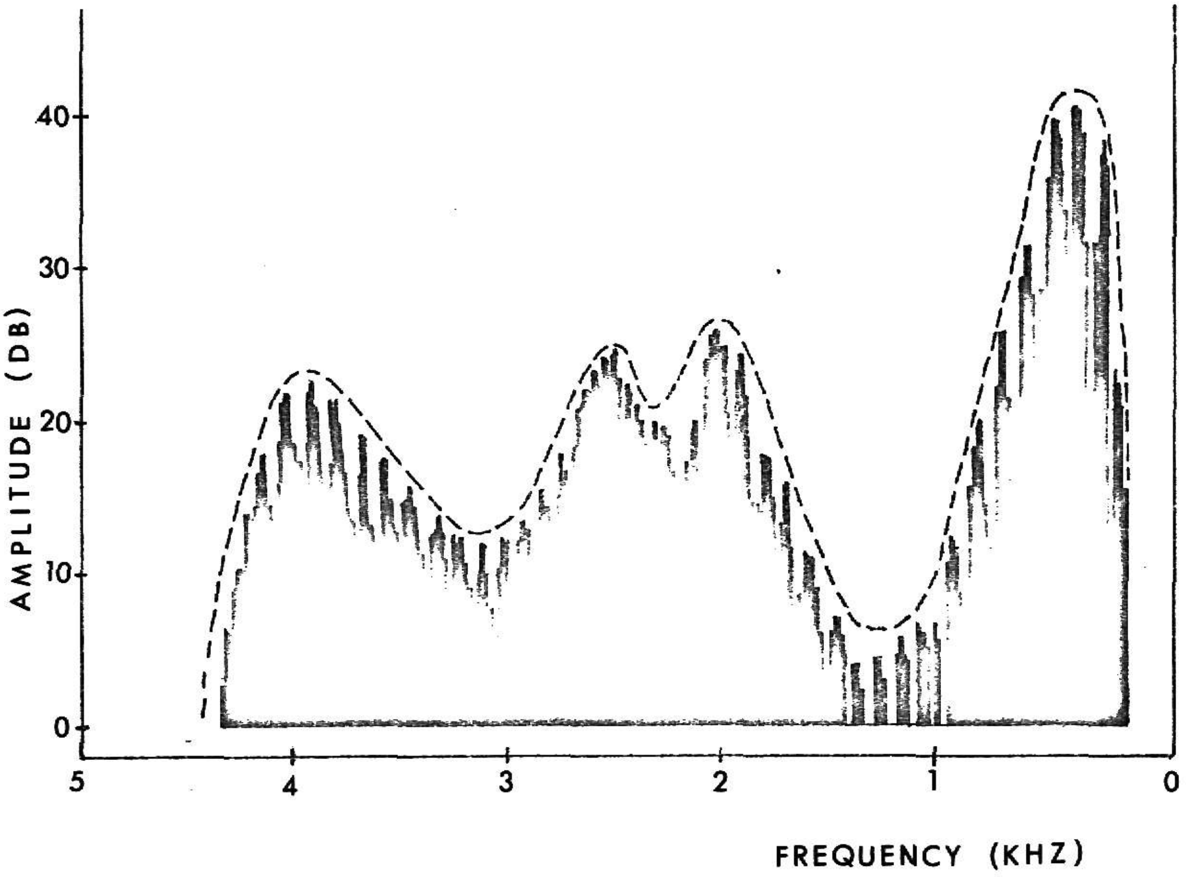

Figure 3.2-4 Logarithmic spectrum of "ee" sound in PROCEEDINGS

much with ongoing time. These imply a monotonous experience for a listener in terms of pitch.

3.3 Articulatory, Voice and Language Related Differences [1, 6]

The three principal areas in which exist considerable difference between the speech of a deaf person and the one of the normal person are (1) articulatory characteristics, (2) aspects of voice, and (3) matters of language.

Articulatory characteristics. There are seven categories of errors in the articulation of consonants and five categories for consonants:

(i) Voiceless-voiced confusions,

(ii) Substitution of one consonant for another,

(iii) Nasality,

(iv) Mishandling of consonantal clusters,

(v) Mishandling of abutting consonants in different syllables,

(vi) Omission of terminal consonants, and

(vii) Distortion or failure in releasing closely confined air pressure.

Categories for the vowels:

(i) Substitution of one vowel for another,

(ii) Improper sequencing within a diphthong,

(iii) Diphthongizing a vowel,

(iv) Nasalizing a vowel, and

(v) Neutralizing a vowel.

Most of the articulatory errors in the speech of the deaf involve incorrect movements of the articulators. In particular, the linking together of successive phonemes is prone to error. It has been shown that [22] deaf speakers generally tend to distort the durational characteristics of phonemes. Again, systematic differences in duration as a function of phonetic environment that are characteristic of normal speech (e.g., vowels are longer when they precede voiced consonants as opposed to voiceless consonants) are generally distorted for deaf speakers [14].

Errors of articulation that do not involve movements of articulators are relatively less severe in that they mainly lead to substitutions rather than unidentifiable inarticulate sounds. In a study [22] , it was found that the formants of the vowels produced by deaf children tend towards values typical of the neutral vowel |a|, as in alone. It was also observed that deaf children tend to use a higher voice frequency than normal-hearing children. Similar results have also been reported by Green [15], Martony [16], and others.

Aspects of voice. The most noticeable feature of deaf speech is the severity of suprasegmental errors. The flow of air through the trachea and larynx of the deaf is poorly coordinated with voice production. This feature affects intonation, which is frequently flat and monotonous. Again, rhythm is caused to be lacking or incorrect. This results in a substantial reduction in intelligibility.

In addition to intonation and rhythm errors, phrasing can also be completely inaccurate.

Deaf speakers as a rule produce sounds inefficiently and have to pause for breath more frequently [17]. They also tend to speak more slowly and the combined effect is to have short breath groups encompassing only a few words at a time. The importance of good respiratory control was emphasized by Hudgins [18], who found a marked reduction in intelligibility due to poor breathing habits. Hudgins found that poor respiratory control leads not only to incorrect groupings of syllables, but also, to improper placing of stress.

Matters of language. The speech communication depends upon a variety of features. One of them is to have an adequate knowledge of the language. This condition counts for persons with normal hearing also. For deaf persons, this knowledge does not develop sufficiently, no matter what kind of education they get.

CHAPTER IV

SOME CONSIDERATIONS OF DIGITAL SIGNAL PROCESSING OF DEAF SPEECH

4.1 Introductory Remarks

Until recently, all of the studies pertaining to the acoustic characteristics of deaf speech have used analog-filter systems (e.g., the sound spectrograph), in obtaining the data. Compared to digital methods, these techniques are relatively imprecise. Digital techniques also offer great flexibility in the choice of filter bands, including the width and shape of the spectral windows that are used. This is important, since there is always a conflict between the requirements of bandwidth and the duration integration time, especially since the time-varying characteristics of deaf speech may differ from those of normal speech. The most significant advantage, however, is that digital techniques provide the investigator the freedom in choosing the most appropriate method of analysis for a given problem.

A second major advantage of computerized techniques lies in its ability to handle large quantities of data. One of the problems encountered in the analysis of deaf speech is the large variability between speakers. Differences between deaf speakers are much greater than differences between normal speakers. Thus, more data is needed to separate out the differences between speakers from characteristic differences between deaf and normal speech.

4.2 Digital Spectrograms

A typical means for obtaining and displaying speech spectra is the spectrograph machine, for which the analysis corresponds to playing the speech through a bank of equal-bandwidth filters. This bank of filters is usually implemented by heterodyning the signal past a single fixed filter. In a narrow-band analysis, the filter bandwidths are typically 45 Hz; for a wide-band analysis, they are 300 Hz. The recording is made on Teledeltox paper. Fig. 4.2-1 shows narrow-band and wide-band spectrograms for the sentence, "He took a walk every morning," as spoken by a male speaker, and reported by Oppenheim.[1] In Fig. 4.2-1(a), it is clear that the individual pitch harmonics have been resolved in frequency, whereas in Fig. 4.2-1(b), they are no longer evident. In the wide-band spectrogram however, vertical striations that correspond to individual pitch periods can be seen. This results from the good time resolution of the wide-band spectrogram; they are not evident in the narrow-band case. Also evident on both wide-band and narrow-band spectrograms are the formants which correspond to high spectral amplitude

[1] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," <u>IEEE Spectrum</u>, August 1970, pp. 57-62.

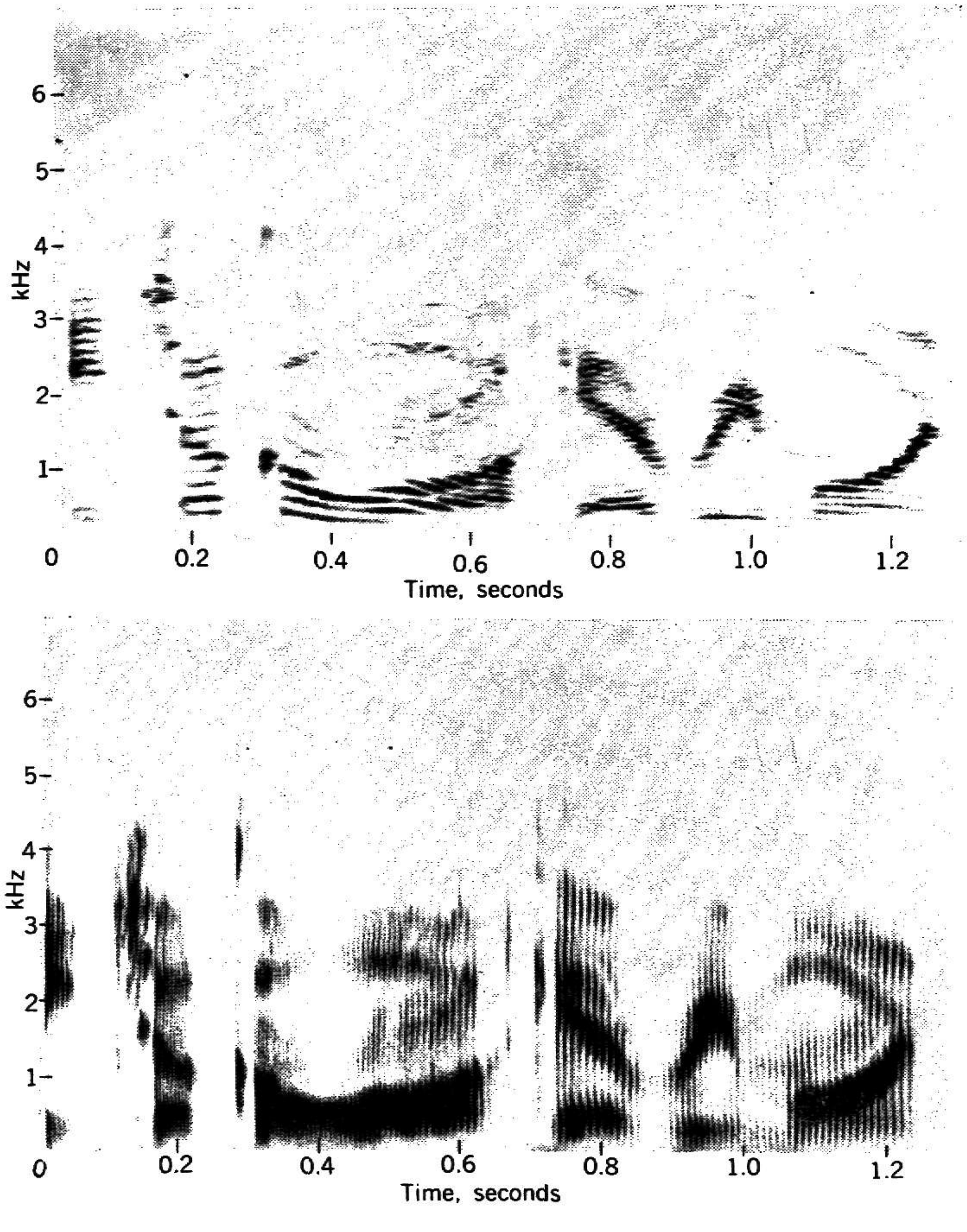

Figure 4.2-1 Narrow-band and wide-band spectrograms for the sentence, "He took a walk every morning"

regions. Such spectrograms can be referred to as analog spectrograms to indicate that they have been obtained via analog filters used in the construction of the spectrograph machine.

In the above study, Oppenheim has also demonstrated that the fast Fourier transform (FFT) algorithm can be used to compute digital spectrograms. The FFT has assumed great importance as a means for computing spectra and implementing spectral displays on a digital computer [19, 20]. The FFT is an algorithm for computing the discrete Fourier transform (DFT), defined as

$$F(k) = \sum_{n=0}^{N-1} f(nT)\, e^{-j\frac{2\pi}{N}nk}, \quad k = 0,1,\ldots,(N-1) \qquad (4.2\text{-}1)$$

where f(nT) corresponds to equally spaced samples of an analog time function f(t). Assuming that the sampling has been done at a rate equal to or higher than the Nyquist rate ($2f_m$, where f_m is the highest frequency in the analog time function), it can be shown that the magnitude of the k-th spectral point $|F(k)|$ in Eq. (4.2-1) corresponds to the magnitude that would be obtained at a time $t = (N-1)T$, when samples of the analog function f(t) are played through an analog filter with a frequency response H(w) given by

$$H(w) = \frac{\sin \frac{NT}{2}\left(w - \frac{2\pi k}{NT}\right)}{\left(w - \frac{2k}{NT}\right)}$$

This filter characteristic is sketched in Fig. 4.2-2 for k = 0. The set of numbers F(k), for k equal to 0 through (N-1), then corresponds to the set of outputs from a filter bank, each filter of which has a spectral shape similar to Fig. 4.2-1, with a center frequency at w = 2πk/NT.

The computation in Eq. (4.2-1) provides only one spectral section, that is, the output of the filter bank at a time t = (N-1)T. To obtain a short-time spectral analysis, it is preferred to perform this computation at successive instants of time, and, in addition, to be able to modify the filter shape. For example, it may be desired to reduce the sidelobes in the filter characteristics in Fig. 4.2-2. Furthermore, as it is changed from a wide-band to a narrow-band analysis, it would be required to change the width of the central lobe in the filter. To determine a running spectrum and provide flexibility in terms of the filter characteristic, the expression in Eq. (4.2-1) can be modified as

$$F_r(k) = \sum_{n=0}^{N-1} w(nT)f(nT+rMT)e^{-j\frac{2\pi}{N}nk} \qquad (4.2\text{-}3)$$

Equation (4.2-3) introduces two changes. The first is to include a window w(nT) to provide a better spectral characteristic. This is motivated by the fact that since a computation of the DFT as given by Eq. (4.2-1) is necessarily restricted to a computation on a finite length of data, there is implicit in Eq. (4.2-1) a time window imposed on f(t); that is, f(t) is multiplied by a rectangular window with a width equal to NT.

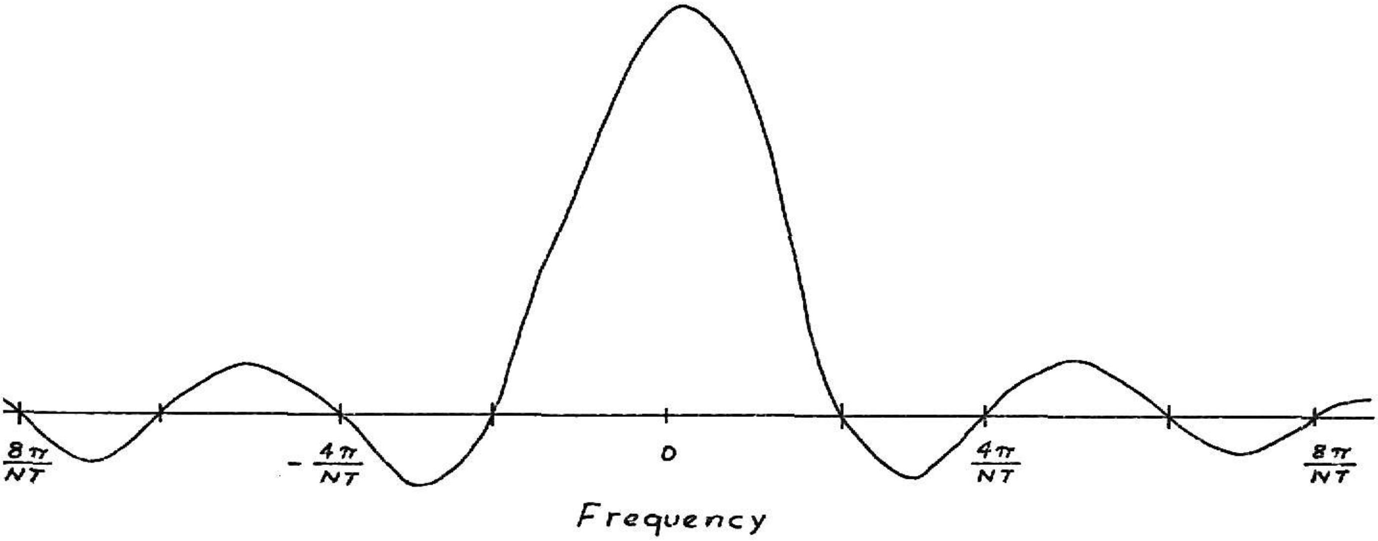

Figure 4.2-2 Equivalent filter characteristic for rectangular time window

It is that rectangular time window that leads to the spectral window shown in Fig. 4.2-2. By modifying the rectantular window with some new time window w(nT), it is possible to modify the spectral shape shown in Fig. 4.2-2. The second modification incorporated in Eq. (4.2-3) corresponds to implementing a spectral analysis of successive sections of the waveform. In other words, the set of numbers $F_r(k)$ represents a computation of the DFT of a section of the analog time function starting at t = rMT and ending at t = rMT+(N-1)T. This corresponds to a filter bank output at time t = rMT+(N-1)T. Successive sections are spaced in time by MT (see Fig. 4.2-3).

In a filter-bank implementation of the spectral analysis, the time window w(nT) corresponds to the low-pass prototype of the impulse response of each of the filters. One observation from this is that the spectral analysis described by Eq. (4.2-3) corresponds to a filter-bank analysis for which the spectral shape of each of the filters in the filter-bank is approximately the same. For example, Eq. (4.2-3) could not represent a filter bank having constant-Q filters, for which the bandwidth is proportional to the frequency. If a constant-Q analysis were desired, a direct implementation of the filters would be used [21].

As an example of speech spectrograms obtained by using the FFT, the procedure was implemented on the Univac 1219 computer facility at the M.I.T. Lincoln Laboratory. This computer, which is similar in size and speed to those in many

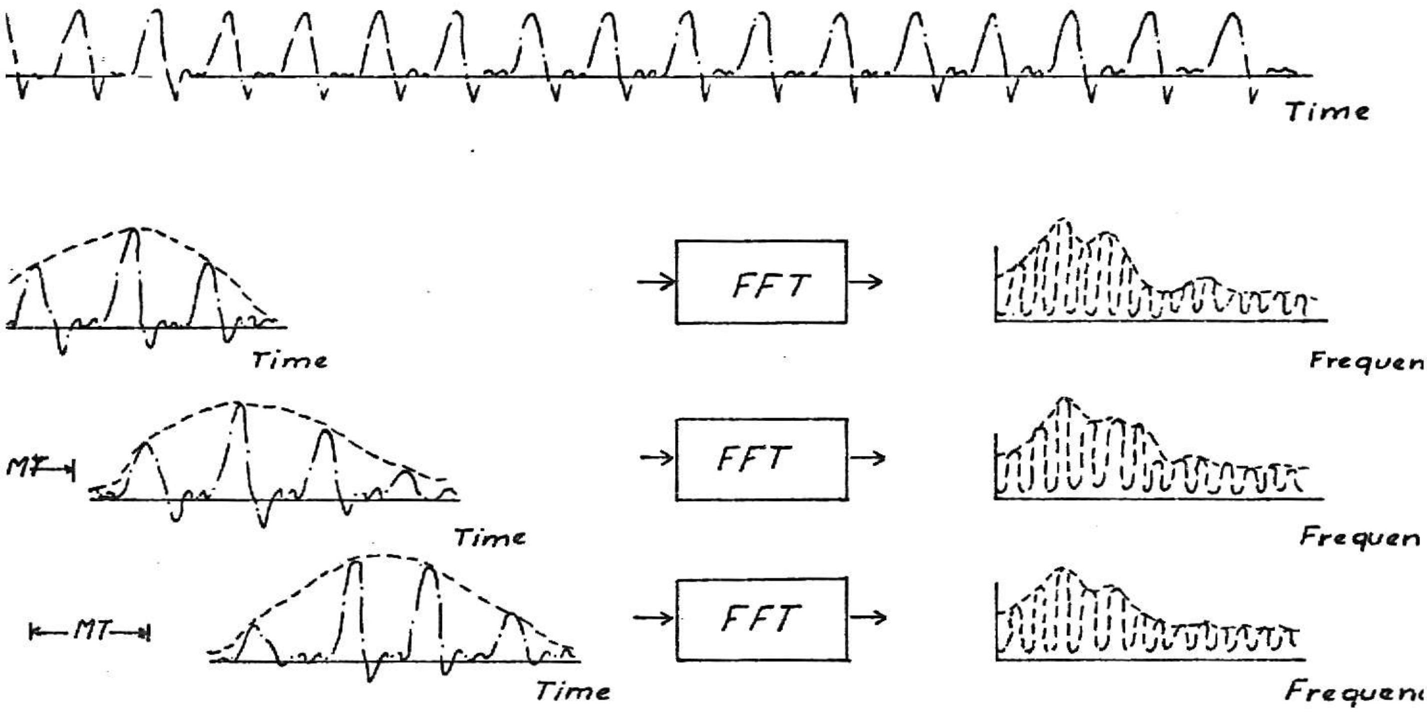

Figure 4.2-3 Computation of running spectrum using FFT

speech laboratories, has a memory cycle time of 2μs, an 18-bit register length, and generally utilizes fixed-point arithmetic. The implementation to be illustrated was programmed in assembly language. One of the objectives was to have an analysis time comparable to that required with modern analog spectrographic equipment. For a narrow-band spectrographic analysis, the number of spectral points computed corresponding to the parameter N in Eq. (4.2-3), is larger than that required for a wide-band analysis. On the other hand, since the time resolution is worse in the narrow-band than in the wide-band case, spectral sections need be computed less frequently for narrow-band spectrograms; that is, the value of the parameter M can be larger. In the system implemented, M was chosen as a fixed percentage of N. Since computation via the FFT algorithm requires a time proportional to $N \log_2 N$ (assuming N is a power of 2), the analysis time for an utterance is essentially the same for either a narrow-band or a wide-band analysis. On this computer the analysis time was approximately three minutes for a two-second utterance. (This will vary somewhat, depending on the speed of the computer.) With the values used for M and N, a sufficient number of points were obtained to define the short-time spectrum with the appropriate time and frequency resolution. Linear interpolation in time and frequency between these samples was utilized to provide a smooth display. A hard copy is obtained photographically, with a time exposure.

The time window w(nT) in Eq. (4.2-3) was chosen to be a Hanning window, defined as

$$w(nT) = \frac{1}{2}\left[1-\cos\frac{2}{NT}nT\right] \qquad 0 \leq n \leq N \qquad (4.2\text{-}4)$$

The corresponding spectral window is shown in Fig. 4.2-4. For both the wide-band and narrow-band analysis, the input speech was pre-emphasized at 6 db per octave starting at 350 Hz,high-pass filtered to 5 KHz, and sampled at 10 KHz. For a wide-band analysis (see Fig. 4.2-5), the parameter N was chosen as 128. This means that the half-power filter bandwidths were approximately 112 Hz and the separation between successive spectral samples was 78 Hz. For the narrow-band analysis (see Fig. 4.2-6), N was chosen as 512 corresponding to half-power filter bandwidths of 28 Hz and a difference between the center frequencies of successive filters of 20 Hz. The parameter M was chosen as 24 for the wide-band analysis and 96 for the narrow-band analysis. This corresponds to obtaining spectral sections every 2.4 ms in the wide-band case and every 9.6 ms in the narrow-band case. The assumption is that values in between can be obtained by interpolation.

It would be advantageous to study the differences between normal and deaf speech using digital spectrograms due to the increased flexibility offered by the latter.

4.3 Formant Estimation Using the Chirp Z-transform [10]

The chirp Z-transform (CZT) is a powerful spectral transformation. This transformation is very appropriate for

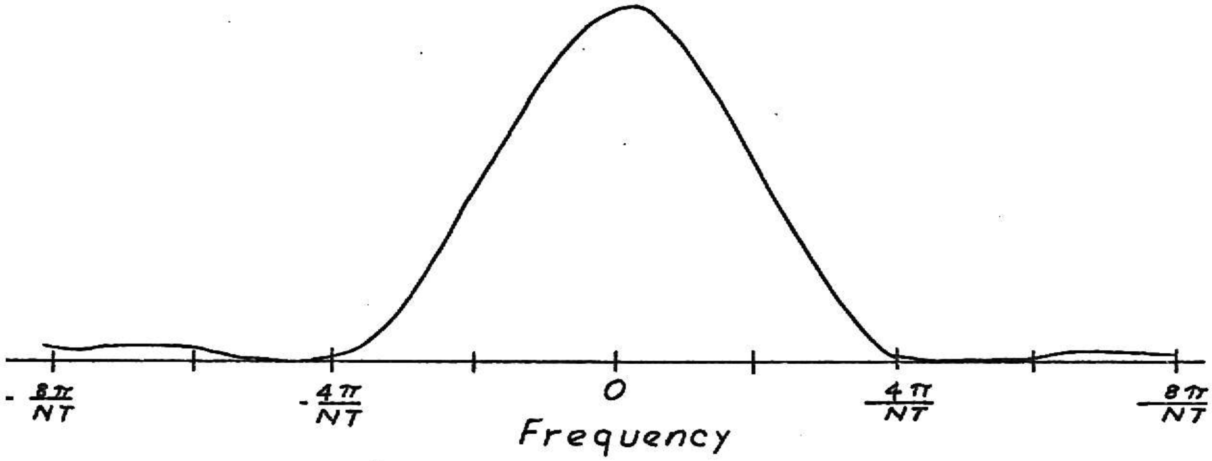

Figure 4.2-4 Equivalent filter characteristic for Hanning window

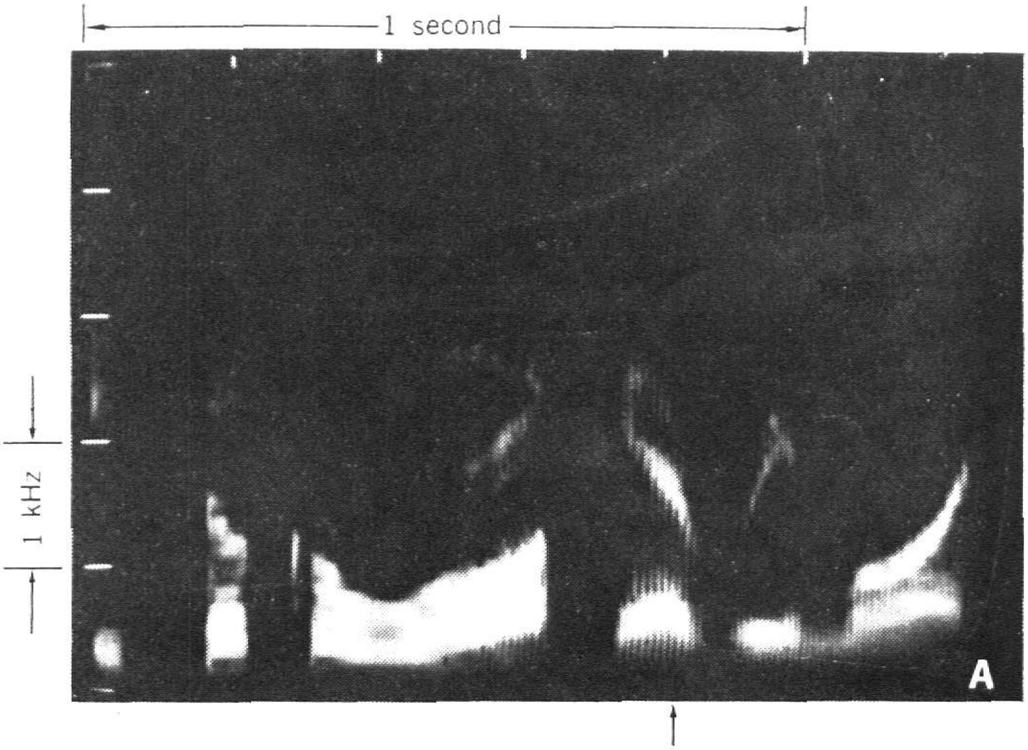

Figure 4.2-5 Wide-band analysis with N chosen as 128

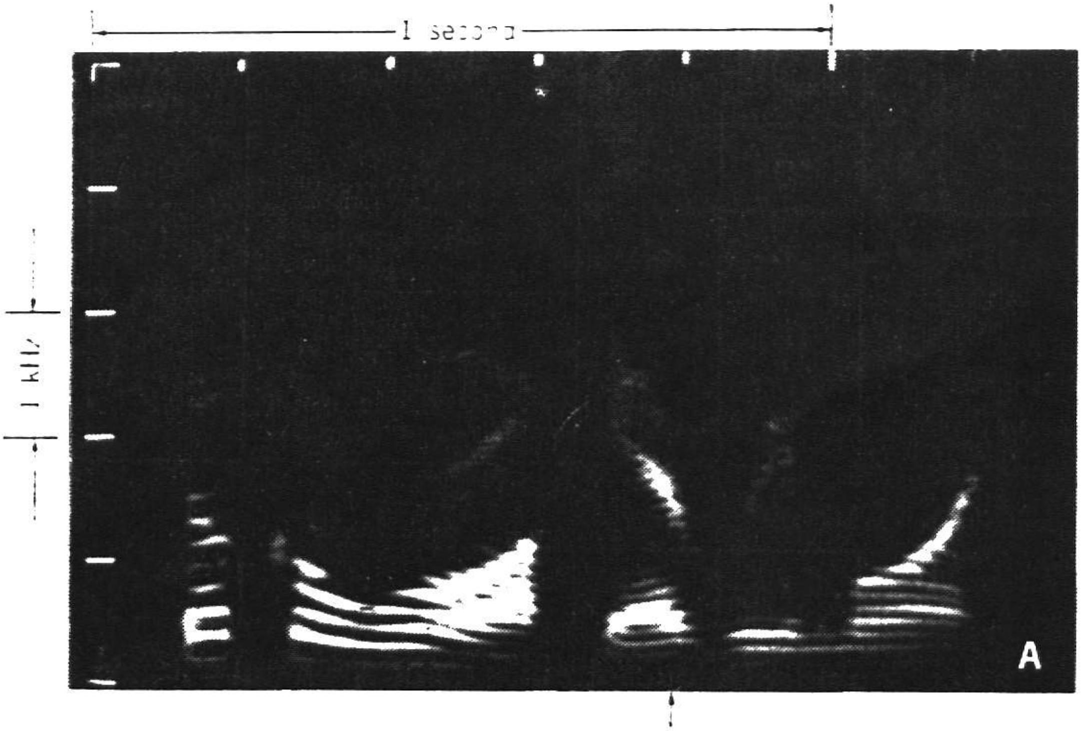

Figure 4.2-6 Narrow-band analysis with N chosen as 512

the estimation of the formant frequencies.

The main difficulty in the formant frequency estimation is to measure the frequency and bandwidths of the resonances. Using the CZT, formant estimates can be obtained in an accurate manner from the pole-zero diagram by evaluating F(s) along a contour that pass close to the poles. This is traditionally accomplished by evaluating F(s) along the jw-axis.

In order to see how the CZT is applied in speech, a specific case considered by Levitt [10] is presented in what follows.

Figure 4.3-1 shows the typical pole locations for the vowel $|i|$. Starting with the pole-locations diagram, the CZT is computed. The CZT is defined as

$$X_k = \sum_{n=0}^{N-1} X_n A^{-n} W^{nk} \qquad k = 0,1,\ldots \quad M-1 \qquad (4.3-2)$$

Where M is an arbitrary integer. A and W are arbitrary complex numbers of the form

$$A = A_o e^{j2\pi\theta_o}$$

$$W = W_o e^{j2\pi\Psi_o}$$

W defines the contour in the s-plane over which the transform is evaluated. A defines the starting point of the contour and M defines the number of points to be evaluated on that contour.

For the above example, the vertical sloping lines are chosen to be the contours of integration. This choice

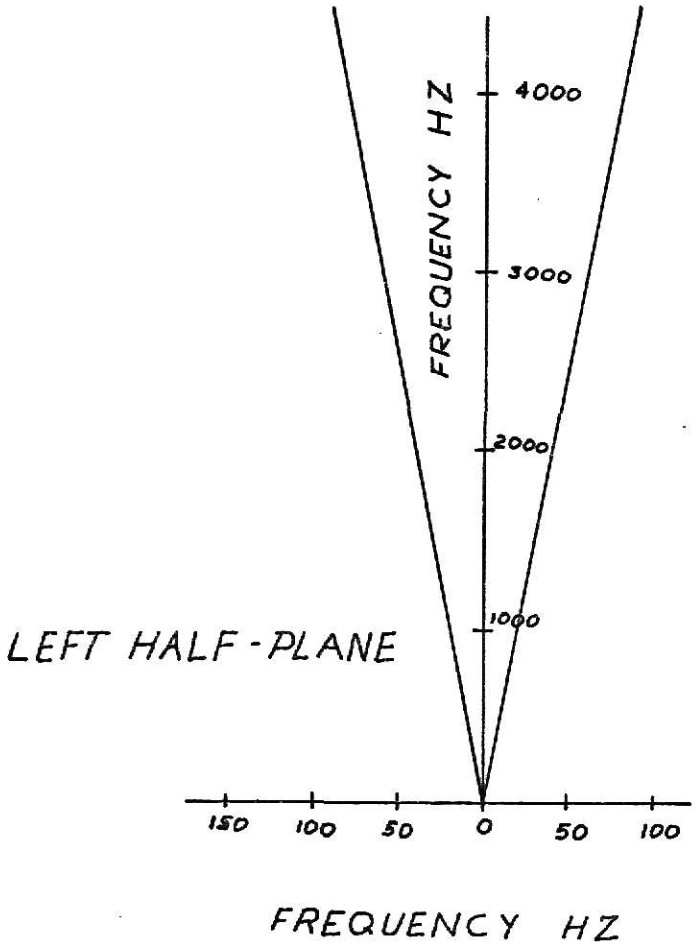

Figure 4.3-1 Pole locations

results in the parameters

$$A = 1$$

$$|W| = e^{-.0004\pi} \text{ for contour 1}$$

$$|W| = 1 \text{ for contour 2 and}$$

$$|W| = e^{.0004\pi} \text{ for contour 3.}$$

The CZT may be computed quite efficiently using the FFT. The results obtained are shown in Fig. 4.3-2. It can be seen that the spectrum obtained with $|W| = 1$ [see Fig. 4.3-2(b)] is the same as the spectrum obtained with traditional techniques. One advantage in using the CZT is thus evident. The resolution with which formant frequencies and bandwidths may be measured is greatly enhanced [see Figs. 4.3-2(a) and 4.3-2(c)].

In the above study, Levitt remarks that the CZT could be particularly useful for estimating deaf speech formants, since conventional spectrograms of profoundly deaf speakers are rather difficult to read. This is because deaf speakers do not move their articulators effectively, and as a result produce sounds with reasonably shady formants, but with marked change in the characteristics of voicing, including some additional noise or friction during the midportion of the utterance.

4.4 Short-term Polynomial Analysis [10]

One of the principal objectives of the acoustic analysis of speech of deaf speakers is to determine how the acoustic parameters of deaf speech differ from those of

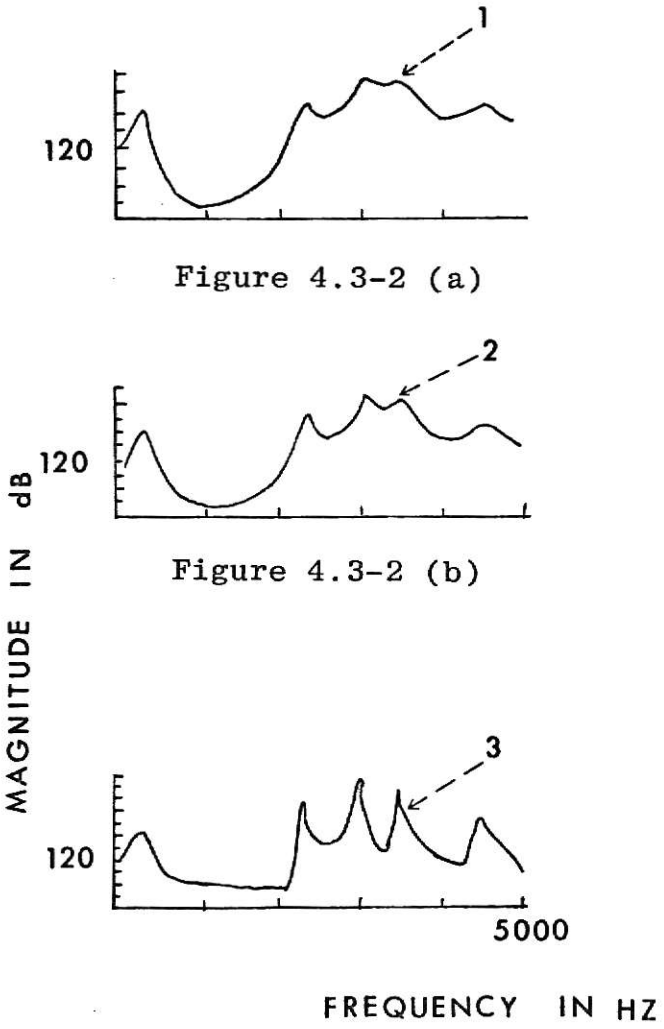

Figure 4.3-2 (a)

Figure 4.3-2 (b)

Figure 4.3-2 (c)

Figure 4.3-2 CZT spectra

normal speech. The problem is made more difficult by large differences between speakers. In addition, such basic data as the fundamental frequency and formant contours are time-varying functions. Whereas standard analysis-of-variance techniques can separate between-speaker and between-group differences for a single parameter, comparing sets of contours presents difficulties. One method is to represent each contour by a set of orthogonal polynomials and subsequently compare coefficients of these polynomials for between-group and within-group differences. Since different portions of these contours may show different variations between and within groups, the orthogonal representation is used over portions of the contour at a time. Hence, the name "short-term polynomial analysis."

Some results of a short-term polynomial analysis of deaf speech for fundamental frequency contours has been reported by Levitt [10]. It has been observed that this type of analysis is equivalent to obtaining on discrete orthogonal transform representation of the formant contours. In particular, the specific short-term polynomial analysis carried out by Levitt was found equivalent to obtaining a BIFORE (Binary Fourier Representation) transform[1]

[1] N. Ahmed, *et al.* "BIFORE or Hadamard Transform," *IEEE Trans. Audio Electroacoust.*, Vol. AU-19, pp. 225-234, Sept. 1971.

representation. The results show that "Moderately" deaf speakers have more erratic formant contours than those of "normal," "Good" and "Poor" speakers. One explanation for this phenomenon is that Moderate speakers made great efforts to articulate each phoneme as best they can with consequent disruptions of the fundamental frequency contour. On the other hand, the poor speakers omit many of the phonemes, particularly the consonants, thereby producing fewer disruptions of the fundamental frequency contour, but with much less intelligible speech.

CHAPTER V

CONCLUSIONS

From the discussion in the previous sections it is clear that in order to study the differences between the acoustic characteristics of normal and deaf speech, there are two essential tools a researcher needs. These are: (i) flexibility for changing parameters associated with the method of analysis, and, (ii) capability of handling large amounts of data. Both these requirements can be satisfied by using digital signal processing techniques.

It is recommended that differences between the acoustic characteristics of normal, slightly deaf, moderately deaf and profoundly deaf speech be studied using digital processing methods. The following methods of analysis may be used:

(i) The CZT [10],

(ii) Oppenheim's digital spectrogram [23],

(iii) Arnold's digital spectrogram [21].

The CZT enables a high resolution frequency analysis of the formant contours. Again, while Oppenheim's spectrogram consists of a bank of digital filters with constant bandwidth (variable Q), Arnold's spectrogram consists of a bank of variable bandwidth (constant Q). For example, Fig. 5.1-1 shows Arnold's spectrogram (i.e., a time-frequency-amplitude plot) of a 4Hz damped sinusoidal waveform, sampled at

Frequency →

Time(sΔt) ↓

s	1.32	1.55	1.74	1.96	2.20	2.46	2.73	3.14	3.53	3.97	4.47	5.03	5.65	6.36	7.15	8.05	9.06	10.19	11.46	12.35
0	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	8	8	8	8	8	8	8	8	8	7	7	7	6	6	5	3	2	0	-5	-9
2	9	9	9	9	9	9	8	8	8	8	7	6	6	5	3	2	0	-2	-6	-9
3	8	8	8	8	8	9	9	9	9	9	9	9	8	7	5	2	-3	-9	-2	0
4	7	8	9	9	10	10	11	11	11	11	11	10	9	6	2	-6	-4	0	-1	-9
5	8	9	9	10	11	11	12	12	12	12	11	10	8	5	0	-3	-2	-3	-5	-5
6	7	8	9	9	10	11	11	12	12	12	12	11	9	5	-1	-3	0	-2	-9	-1
7	5	6	7	8	9	10	11	12	13	13	13	11	8	2	-2	1	-1	-8	-1	-9
8	4	5	6	7	9	10	12	13	13	13	13	11	7	2	0	0	-3	-3	-5	-4
9	4	5	6	7	9	10	12	13	13	13	13	11	7	1	0	0	-5	-2	-6	-3
10	6	6	6	7	8	9	11	12	13	14	13	11	6	1	2	-1	-2	-3	-3	-8
11	6	6	6	6	7	9	11	13	14	14	13	10	6	3	1	-2	-1	-5	-4	-4
12	6	6	6	6	7	9	11	13	14	14	13	10	6	3	1	-2	-1	-5	-4	-4
13	6	6	7	7	8	9	11	12	14	14	13	10	5	4	1	0	-3	-2	-6	-7
14	6	7	7	7	8	8	10	12	14	15	13	10	5	4	0	0	-3	-4	-4	-4
15	6	7	7	7	8	8	10	12	14	15	13	10	5	4	0	0	-3	-4	-4	-5
16	6	7	7	7	8	9	10	12	14	15	14	10	6	4	0	0	-2	-4	-6	-6
17	6	6	7	8	8	9	10	12	14	15	13	9	6	3	1	-1	-2	-3	-4	-4
18	6	6	7	8	8	9	10	12	14	15	13	9	6	3	1	-1	-2	-4	-5	-5
19	6	6	7	8	8	9	10	12	14	15	13	9	6	3	1	-1	-3	-4	-5	-5
20	6	6	7	7	8	9	10	12	14	15	13	9	6	3	1	0	-2	-4	-5	-5
21	6	6	7	7	8	9	10	12	14	15	13	9	6	3	1	0	-2	-4	-5	-5
22	6	6	7	7	8	9	10	12	14	15	13	9	6	3	0	0	-2	-3	-4	-5
23	6	6	7	7	8	9	10	12	14	15	13	9	6	3	0	-1	-2	-4	-5	-5
24	6	6	7	7	8	9	10	12	14	15	13	9	6	3	0	-1	-2	-4	-5	-5
25	6	6	7	7	8	9	10	12	14	15	13	9	6	3	0	-1	-2	-4	-5	-5

Figure 5.1-1 Arnold's spectrogram of a 4HZ damped sinusoidal waveform

26 samples/second; that is, $\Delta t = 1/26$ sec. The spectrogram has been computed as a set of 20 specified frequencies which are equally spaced on a logarithmic scale, and given by

$$\{f_k\} = \{ 1.32, 1.55, 1.74, 1.96, 2.20, 2.48, 2.79, 3.14, 3.53, 3.97, 4.47, 5.03, 5.65, 6.36, 7.15, 8.05, 9.06, 10.19, 11.46, 12.89\}.$$

The filter outputs have been scaled by a convenient scale factor and subsequently converted to db. The db. value so obtained is denoted by $d(f_k, s\Delta t)$. The manner in which the plot can be interpreted is best illustrated by the following examples:

(i) $d(3.97,15) = 15$ and $d(1.96,2) = 9$ implies that the power in the 3.97 Hz filter is 6 db. greater than that in the 1.96 Hz filter.

(ii) $d(2.48,6) = 11$ and $d(10.19,6) = -2$ implies that the power in the 2.48 Hz filter is 13 db. greater than that in the 10.19 Hz filter.

It is recommended that such plots obtained by analyzing phonetically balanced sentences such as "I like happy movies," "Noon is a happy time of day," etc. Subsequently these plots can be used to study the differences between acoustic characteristics. Finally, the differences so obtained can be correlated with corresponding differences in the speech production mechanisms.

REFERENCES

1. Connor, Leo E. *Speech for the Deaf Child: Knowledge and Use*. Washington, D. C.: Alexander Graham Bell Ass. for the Deaf, 1971.

2. Fant, G. M. *Acoustic Theory of Speech Production*. Production Gravenhage, Netherlands: Mouton, 1960.

3. Stevens, K. N., House, A. S. "An Acoustical Theory of Vowel Production and Some of its Implications," Journal of speech and hearing research, 1961, 4, 303.

4. Heinz, J. M., Stevens, K. N. "On the Properties of Voiceless Fricative Consonant," Journal of the Acoustical Society of America, 1916, 33, 589.

5. Fujimura, O. "Analysis of Nasal Consonants," Journal of the Acoustical Society of America, 1962, 34, 1865.

6. Pickett, James M. "Status of Speech Analyzing Communication Aids for the Deaf," IEEE Transactions on Audio and Electroacoustics, Vol. AU-20, No. 1.

7. Koenig, W., Dunn, H. K., Lacy, L. Y. "The Sound Spectrograph," Journal of the Aoustical Society of America, 1946, 17, 19.

8. Schroeder, M. R. "Vocoders: Analysis and Synthesis of Speech," Proceedings of the IEEE, 1966, 54, 720.

9. Schafer, Ronald W. "A Survey of Digital Speech Processing Techniques," IEEE Transactions on Audio and Electroacoustics, Vo. AU-20, No. 1.

10. Levitt, Harry. "Acoustic Analysis of Deaf Speech Using Digital Processing Techniques," IEEE Transactions on Audio and Electroacoustic, Vol. AU-20, No. 1.

11. Chiba, T., Kajiyama, M. The Vowel--Its Nature and Structure, (Tokyo, 1941).

12. Fano, R. M. "Short-time Autocorrelation Functions and Power Spectra," J. Acoust. Soc. Am., Vol. 22, pp. 546-550, 1950.

13. Schroeder, M. R., Atal, B. S. "Generalized Short-time Power Spectra and Autocorrelation Functions," J. Acoust. Soc. Am., Vol. 34, pp. 1679-1683, November 1962.

14. Calvert, D. R. "An Approach to the Study of Deaf Speech," Report of the Proceedings of the International Congress on Education of the Deaf. Washington, D. C.: U. S. Government Printing Office, 1964, 261-267.

15. Green, D. S. "Fundamental Frequency Characteristics of the Speech of Profoundly Deaf Individuals," Ph. D. dissertation, Purdue University, Lafayette, Indiana, 1956.

16. Martony, J. "On the Correction of Voice Pitch Level for Severely Hard-of-hearing Subjects," Amer. Ann. Deaf, Vol. 113, pp. 195-202, 1968.

17. Levitt, H. "Speech production and the Deaf Child" in Speech for the Deaf Child, L. Connor, ed. Washington, D. C.: A. G. Bell Ass. for the Deaf, 1971.

18. Hudgings, C. V., Numbers, F. C. "An Investigation of the Intelligibility of Speech of the Deaf," Genet. Psych. Mon., Vo. 25.

19. Cooley, J. W., Tukey, J. W. "An Algorithm for the Machine Computation of Complex Fourier Series," Math. Comput., Vol. 29, pp. 297-301, April 1965.

20. Rothavser, E., Mainald, D. "Digitalized Sound Spectrography Using FFT and Multi-print Techniques (abstract)," J. Acoust. Soc. Am., Vo. 45, p. 308, 1969

21. Arnold, C. E. "Spectral Estimation for Transient Wareforms," presented at IEEE Arden House Workshop, January 1970.

22. Angelocci, A., Kopp, G., and Holbrook , A. "The vowel formants of deaf and normal hearing 11 to 14 year-old boys," Journal of Speech and Hearing Disorders, 1964, 29, 156-170.

23. Oppenheim, A. V. "Speech Spectrograms using the Fast Fourier Transform," IEEE Spectrum, August 1970, pp. 57-62.

24. Boothroyd, A. "Acoustics of Speech" in Speech for the Deaf Child, L. Connor, ed. Washington, D. C.: A. 5. Bell Ass. for the Deaf, 1971.

ACKNOWLEDGEMENTS

I want to express my sincere appreciation to my major professor, Dr. Nasir Ahmed, for his valuable assistance in the preparation of this report and for his interest, patience and encouragement throughout my study at Kansas State University.

My appreciation is also expressed to Dr. W. W. Koepsel, Head of the Electrical Engineering Department, and Dr. R. F. Kruh, Dean of Graduate School, for their assistance.

I would also like to thank Dr. H. R. Rainbolt and Dr. D. H. Hummels for serving on my committee.

To my husband, my heartfelt thanks for his patience and kind understanding.

SOME CONSIDERATIONS OF
DEAF SPEECH

by

M. HELENA VERGARA NOLAN

B. S., Universidad Javeriana, Bogota, Colombia, 1971

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1974

The main purpose of this report is to initiate an interdisciplinary effort between the Departments of Electrical Engineering and Speech at Kansas State University, in the area of speech processing of deaf speakers. To this end, this report presents a survey of the pertinent literature. This survey is presented in terms of the differences between the acoustical characteristics of normal and deaf speech, and the articulatory, voice and language related differences.

It is found almost all the efforts related to the study of acoustic characteristics have relied heavily on analog filter systems such as the conventional sound spectrograph, which are relatively imprecise compared to digital methods of analysis. Again, digital methods offer a great deal of flexibility for changing parameters associated with the method of analysis, and are capable of handling large amounts of data. In the case of deaf speech, large amounts of data are encountered due to the large variability between speakers. Thus, a summary of the methods available for the digital processing of speech are presented. In conclusion, recommendations for future research work are included.