AN UPPER BOUND FOR TACTILE RECOGNITION OF SPEECH

by

Richard P. Mc Clellan Jr.

A Thesis Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

1967

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his judgment the proposed use of the material is in the interest of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Arthur L. McClellan Jr.

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

F. J. Hill

F. J. HILL
Associate Professor of Electrical Engineering

31 May 1967
Date

ACKNOWLEDGEMENTS

The author is particularly grateful to Dr. Fredrick J. Hill whose guidance during the research leading to this thesis and advice during preparation of the initial transcripts was invaluable. Sincere thanks must also be extended to Mr. Lorin P. Mc Rae for his assistance and guidance throughout the project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES -- Continued

ABSTRACT

A method for enabling deaf people to recognize speech was discussed. It was proposed that this be accomplished by replacing the cochlea by a speech processor, and the auditory nerve by the tactile nerves. The speech processor was designed from considerations of the characteristics and production of speech. A flexible parallel filter speech processor was designed and built using active R-C filters. An upper bound on tactile recognition of speech from this speech processor output was determined to be the intelligibility of synthetic speech created from the speech processor output. Effects of amplitude quantization upon speech processor output were investigated. A serial filter speech processor designed as an analog to the human cochlea was discussed.

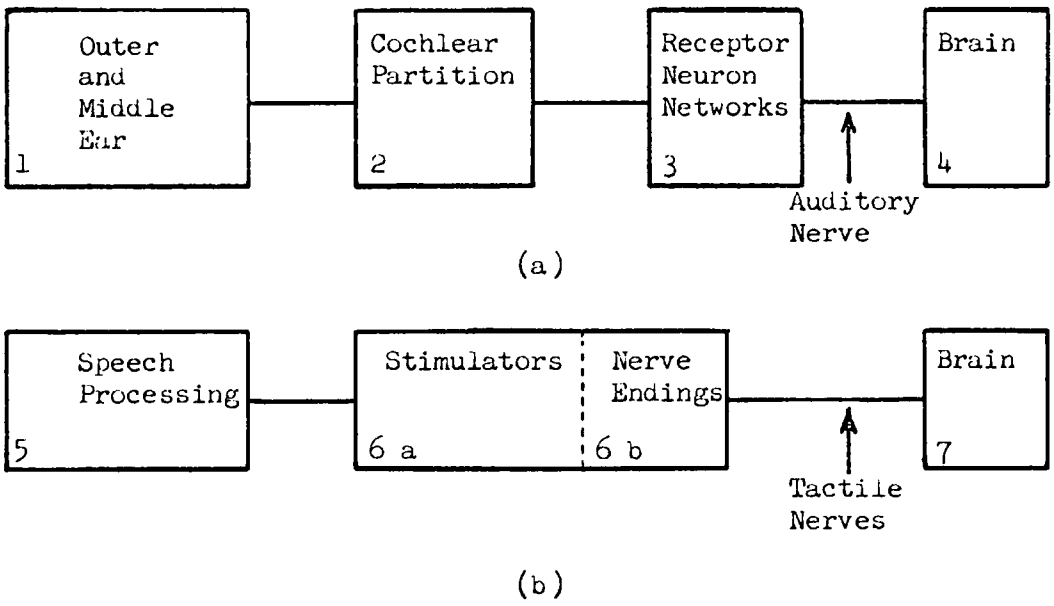
Chapter I

Introduction

The normal person receives information through the senses of sight, smell, taste, feel, and hearing. When one of these senses is lost, the individual compensates by using the remaining senses to a greater extent. The deaf are trained in lip reading and sign language. This substitution, while of great value, has certain disadvantages and limitations. In lip reading, the deaf person must be able to see the talker. When sign language is used, both the "talker" and the "listener" must be trained. In both cases, visual attention is required of the listener.

It would be of great advantage to the deaf if the tactile sense could be substituted for hearing. The deaf person would be able to converse with a person whom he could not see. Television, radio, and the telephone could be used by the deaf.

In order to substitute the tactile sense for hearing, speech information must be introduced into the nervous system by tactile stimulation. A grossly simplified model of the ear appears in Figure 1-1 (a), and a model of a device for tactual introduction of speech information in Figure 1-1 (b). The function of blocks 1, 2, 3 in Figure 1-1 (a) is being supplanted by blocks 5 and 6 (a) in Figure 1-1 (b). Hence, blocks 5 and 6 (a) are, in a sense, an artificial ear.



Simplified Models of Human and Artificial Ear

Figure 1-1

In natural, unprocessed speech, an indication of the speakers emotional state is contained in his pitch inflections. His sex can be determined from his pitch frequency, and the actual message he intends to convey from the words and sentences that he speaks. One of the disadvantages of many speech synthesis schemes, such as Dudley's Vocoder,[1] is that the synthesized speech sounds artificial,[2] due to small variations in pitch frequency being lost. As nothing can now be learned from pitch inflections, information has been lost. However, this synthesized speech is quite recognizable. A measure of the recognizability of speech is the percentage correct recognition of spoken sounds. These sounds may be single phonemes, syllables, words, or sentences.

It has been found that if words are recognized correctly 60% of the time, sentences will be recognized correctly about 95% of the time.[3]

In this study, it was desired to eliminate context as much as possible, so the percentage correct recognition of phonemes, called intelligibility, was taken as the measure of speech recognizability. If the speaker pronounces S phonemes and the listener correctly identified R of these phonemes, then the intelligibility, I, is

$$I = (R/S) \times 100\%.$$

This definition of intelligibility applies nicely to audible speech. Often, speech will be processed into a set of signals which are not audible speech, but from which part of the message contained in the original speech can be extracted. If the speech message is a list of phonemes, and an ideal algorithm for extracting this message were available, the percentage correct recognition of these phonemes by the use of this algorithm is the potential intelligibility of the signals. If P.I. is the potential intelligibility of a signal, S_I is the number of phonemes spoken, and R_I is the number of phonemes correctly identified by the use of the ideal algorithm, then

$$P.I. = (R_I/S_I) \times 100\%.$$

The definition of intelligibility will be extended to any actual speech recognition process. For example the "intelligibility" of speech by the use of any speech recognition technique is the percentage correct phoneme recognition using this technique. Note that the potential intelligibility of any signal is an upper bound on the intelligibility which can be realized.

In Figure 1-1 (b), block 5 processes speech into a form suitable for introduction into the nervous system by the stimulator-nerve ending interface in block 6 (a)-6 (b). Then this speech information is transmitted to the brain on the tactile nerves. The brain recognizes speech from this speech information.

The success of this scheme depends upon several parameters. First, the brain must be capable of processing tactually obtained information at a rate rapid enough to recognize speech. Assuming that the brain is so capable, then sufficient information must be delivered by the tactual nerves.

The success of Braille reading indicates that the brain is capable of processing tactually obtained information at a rapid rate. Whether sufficient information can be delivered by the tactual nervous system is not known at this time. Obviously, no more information can reach the brain than is present at the output of the speech processor. Hence, an upper bound on the intelligibility of tactually communicated speech is the potential intelligibility of the output of the speech processor.

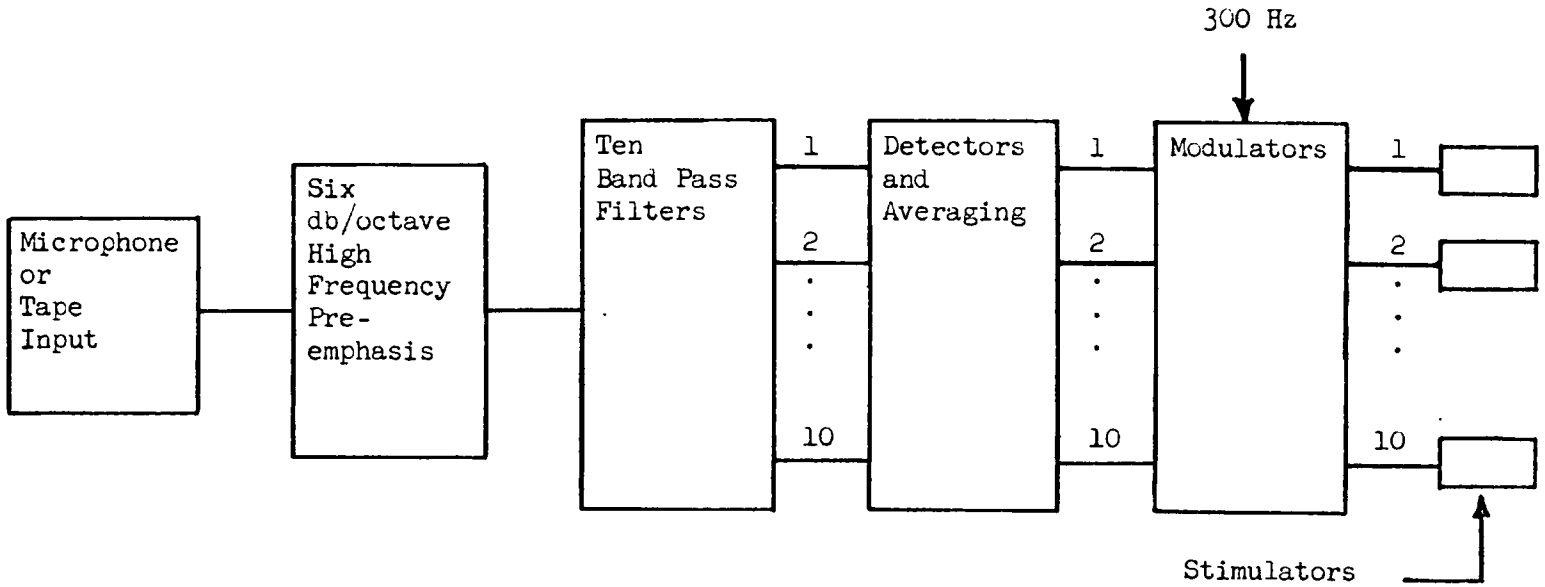
Assuming feasibility of such a device, considerable training will probably be required in its use. How extensive this training will be is not known. Perhaps one of the most significant potentialities of such a device would be its use by a deaf child. If such a child could be introduced to this device at an early age, perhaps in time it would replace the normal hearing apparatus.

The results of several research projects previous to this one indicate the possibility of such a device. Of particular interest is some work done by J. M. Pickett, at the Speech Transmission Laboratory of the Royal Institute of Technology, Stockhold, Sweden.[4] The tactual vocoder shown in Figure 1-2 was used to aid in lip reading.

The bank of filters shown in block 3 covered the range from 210 to 7700 Hertz in ten overlapping ranges. The output of each of these filters was then detected and averaged. This averaged output was modulated with a 300 Hertz sine wave, and used to drive a stimulator placed on one of the ten fingers.

Dr. Pickett found that trained subjects, aided by the tactual vocoder, recognized, by lip reading, 85.5% of selected phonemes spoken by one person. Without the tactual vocoder, recognition fell to 60.2%. For a different speaker, 75.4% of the selected phonemes were recognized with tactual inputs and 66.7% were recognized correctly without. These experiments indicated that information was definitely being transmitted by tactile stimili.

Further experiments were conducted by Dr. Pickett, at the Royal Institute of Technology,[5] on recognition of speech by the use of the tactual vocoder alone. For one set of six vowels, the correct response was made 60% of the time, while with another set of six vowels the correct response was made 45.5% of the time. These tests were run with two speakers, which were also the two subjects tested. It was noted that the vowels with more widely spread formants were more



Pickett's Tactual Vocoder

Figure 1-2

easily distinguished than those with closer spacing. Testing was also done on consonants and it was found that the pair /s-t/ was discriminated correctly 99.5% of the time, while the pair /p-b/ was correctly discriminated only 22.0% of the time. These percentages were corrected for chance success.

Dr. Pickett concluded that the tactual nervous system is capable of transmitting limited amounts of information, and also that the limitations of this method will be imposed primarily by the pattern resolving capability of the skin. Therefore, he feels that the primary applications of tactual communication of speech will be either in situations where severe limitations are placed on normal speech communication, or as a complement to some other information input.

For example, a limited vocabulary could be tactually recognized by a person working in a high noise environment where normal voice communication would be impossible. Dr. Pickett's original application of the tactual vocoder was a supplement to lip reading, where it proved to be of value.

Dr. Pickett's tactual vocoder falls into the pattern of Figure 1-1 (b). The tactual vocoder used a bank of ten filters and envelope detectors as the speech processing element. As previously outlined, the potential intelligibility of the speech processor output must be sufficiently high, if satisfactory operation is to be realized. It is intuitively reasonable that in a parallel filter speech processor such as this, the potential intelligibility of the output increases as

the number of channels increase. Since this device is to be reasonably compact, it is desirable to use as few channels as necessary.

It is not immediately obvious what the minimum number of channels is for adequate potential intelligibility. Dr. Pickett used ten channels, presumably with the idea in mind of using the ten fingers as receptors. Other investigators have used as many as one hundred and as few as eight channels,[6] but no one has determined how potential intelligibility varies with the number of channels. Indeed, it is a difficult problem to attack directly as potential intelligibility is not easily measured. Also, it is reasonable that the manner in which the channels are arranged, as well as the number of channels, is important.

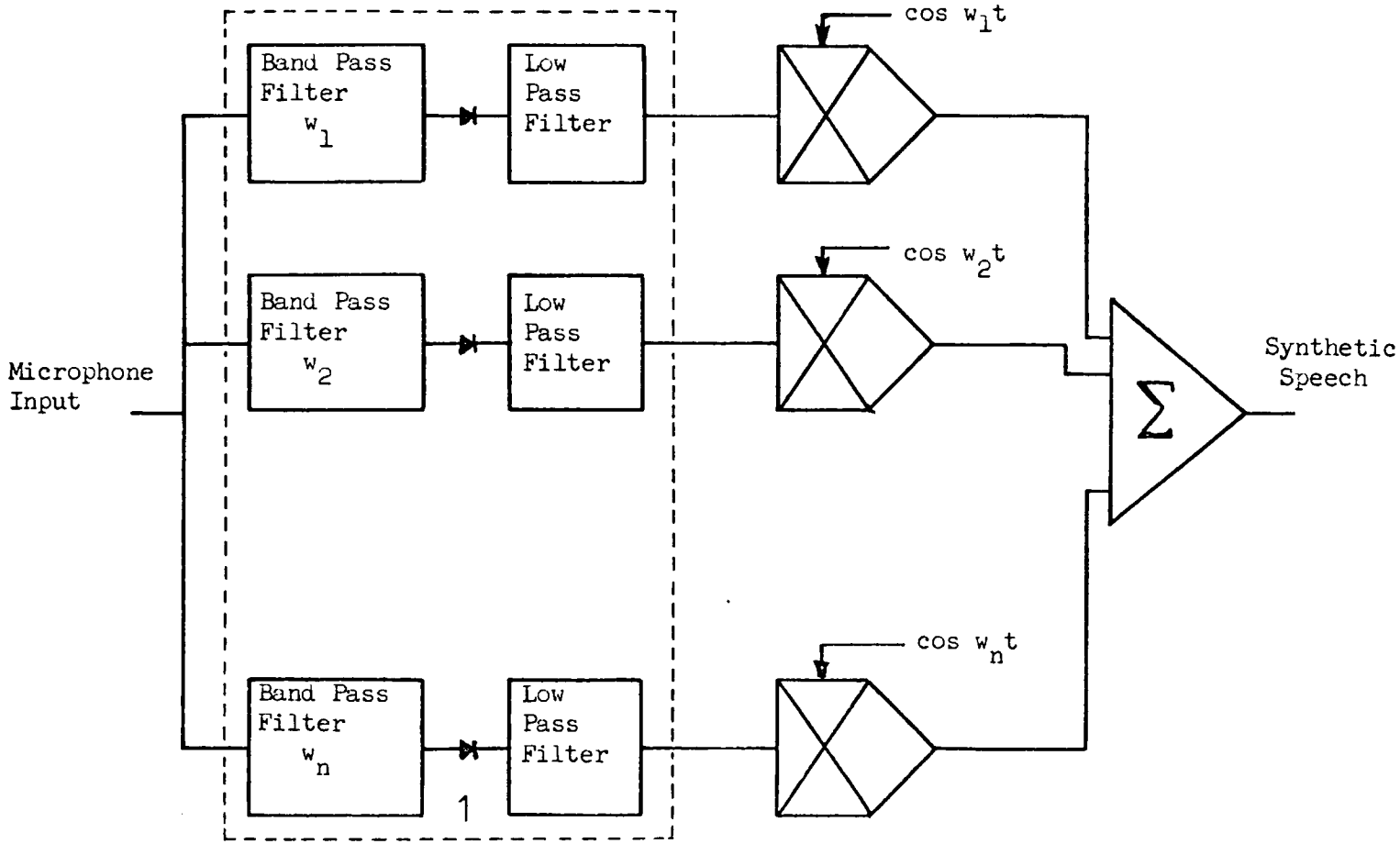
The object of the research being reported was to indirectly obtain data on how potential intelligibility varies with number of channels. One method of accomplishing this would involve building many tactual vocoders, each with a different number of channels. A large number of subjects would be trained in the use of each vocoder. Then the average of these subjects phoneme recognition rates could be used as an indication of the potential intelligibility of the output of that particular speech processor. This approach is impractical due to the long time which would be required for the subjects to learn to recognize speech from tactile stimulation.

Hence, a more direct method of obtaining a measure of the potential intelligibility of the output of the parallel filter speech

processor is needed. If a pattern recognition system capable of recognizing speech with little or no retraining were available, and the capability existed for rapidly changing speech processor configurations, equivalent results could be obtained much more easily.

It might be observed that the auditory system is already such a pattern recognition system. Hence, if some method of utilizing the auditory system to recognize speech from the output of the speech processor were used, the percentage recognition of phonemes would serve as a measure of the potential intelligibility of the output of the speech processor. Figure 1-3 is a block diagram of the system used.

In Figure 1-3, block 1 is the speech processor, the outputs are the low frequency envelope of the outputs of each band pass filter. Each output is used to modulate a sinusoid whose frequency is the center frequency of the respective band pass filter. In this manner, the signal stimulates a region of the basilar membrane rather than a region on the skin. This device is conceptually similar to the vocoder originated by H. Dudley,[7] and the output bears a strong resemblance to speech. Hence, the pattern recognition system in the brain needs little retraining, and the phoneme recognition rate for a particular speech processor configuration can be evaluated in a reasonably short time. It can then be said that the recognition rate observed in this manner will provide an upper bound for tactual recognition using this speech processor.



Speech Processor--Speech Synthesizer

Figure 1-3

Chapter II

Speech Processor Design

The speech processor used in these experiments was the bank of filters similar to that used in Dr. Pickett's tactual vocoder. Referring to Figure 1-1, it is seen that the function of the speech processor is to convert speech information into a form adequate to drive tactile stimulators. The tactile nerve endings are not sensitive to the wide range of frequencies and amplitudes that the auditory system is capable of discriminating.

A closer look at speech may serve as a guide to the design of the speech processor. As the output of the speech processor described in Chapter I depends upon the frequency content of the input, Fourier analysis may prove of interest. The basic linguistic unit is the phoneme. If we represent a single phoneme as a function of time, f(t), then, f(t) will be of finite amplitude and time duration. Then, since f(t) is bounded over a finite interval and zero elsewhere,

$$\int_{-\infty}^{\infty} |f(t)|\, dt$$

exists.[8] Hence, the Fourier Transform

$$F(jw) = \int_{-\infty}^{\infty} f(t)\, e^{-jwt}\, dt$$

exists, where

$$f(t) = \frac{1}{2\pi}\, PV \int F(jw)\, e^{jwt}\, dw.$$

Now,

$$F(jw) = \int_o^\infty f(-t)\, e^{jwt}\, dt + \int_o^\infty f(t)\, e^{-jwt}\, dt$$

if $f(t) = f_e(t) + f_o(t)$

where $f_e(-t) = f_e(t); \quad f_o(t) = -f_o(-t)$

then $f(-t) = f_e(t) - f_e(t).$

Then,
$$F(jw) = \int_o^\infty \left[f_e(t) - f_o(t)\right] \left[\cos wt + j \sin wt\right] dt + \int_o^\infty \left[f_e(t) + f_o(t)\right] (\cos wt - j \sin wt)\, dt$$

$$= 2 \int_o^\infty f_e(t) \cos wt\, dt - j2 \int_o^\infty f_o(t) \sin wt\, dt.$$

So, if

$$A_1 = 2 \int_o^\infty f_e(t) \cos wt\, dt$$

$$A_2 = -2 \int_o^\infty f_o(t) \sin wt\, dt$$

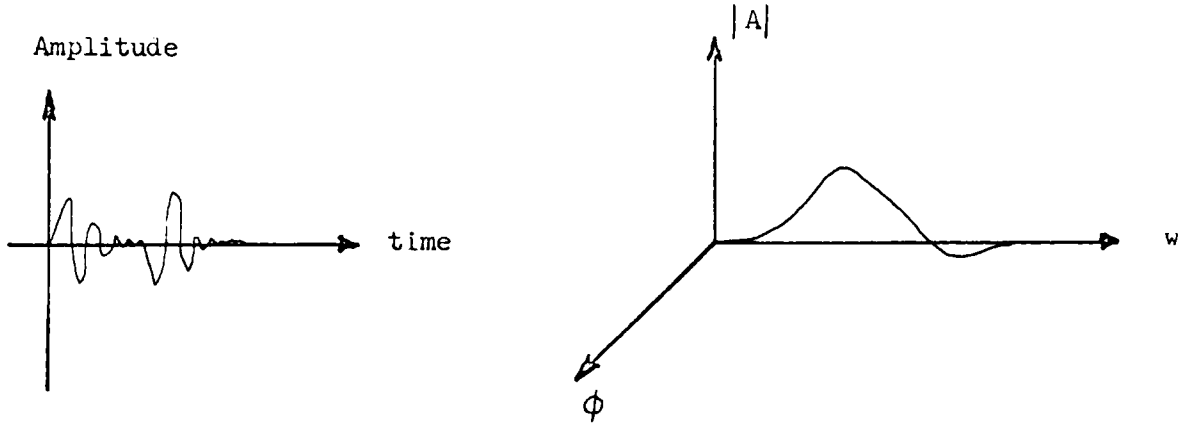
then, $F(jw) = A_1(jw) + j\, A_2(jw)$

$$= \left|A(jw)\right| \underline{\left|\phi\,(jw)\right.}$$

Hence, the phoneme shown as a function of time in Figure 2-1 (a) transforms to a curve in the A, W, ϕ space shown in Figure 2-1 (b).

The exact representation of this curve requires an infinite number of points. This implies that the speech processor in Figure 1-1 (a) has an infinite number of channels. As this is clearly impractical, the question arises as to what constitutes a sufficient approximation. It is reasonable to expect that very high and very low frequencies

contribute little to intelligibility. Fletcher found that speech sounds were recognized correctly essentially 100% of the time if all frequencies below eight KHz were used. However, the frequencies between 250 Hz and 6 KHz contribute over 95% to intelligibility.[9] Hence, if we consider only the portion of the curve within these limits, little potential intelligibility should be lost.



(a) Phoneme (b) Fourier Transform

Fourier Transform of Phoneme

Figure 2-1

At this point it might be instructive to consider further the parameters by which this curve is being described--amplitude, phase, frequency. Exact knowledge of all of these is necessary for precise reproduction of the phoneme. Precise reproduction is necessary for natural sounding speech, or for speech of high quality. However, speech does not have to be of high quality to be quite intelligible. As the object here is speech recognition, quality will be considered of secondary importance.

Manley[10] has found that elimination of relative phase information has a marked effect on speech quality, but little on intelligibility. Since potential intelligibility is of prime interest, it is assumed that the phoneme can be described sufficiently without taking into consideration the phase parameter. Hence, the phoneme will be approximated by a curve in the amplitude frequency plane.

Even with this simplified description, an infinite number of points are required to completely specify this curve. However, if a sufficient number of points on this curve are known, it may be specified completely enough for pattern recognition purposes. It might be noted at this point that the position of the points on this curve as well as the number of points contribute to the accuracy of the approximation.

The preceding discussion is a better theoretical description of a phoneme than guide to the design of block 5, Figure 1-1. Speech does not lend itself well to conventional Fourier analysis. If f(t) represents continuous speech,

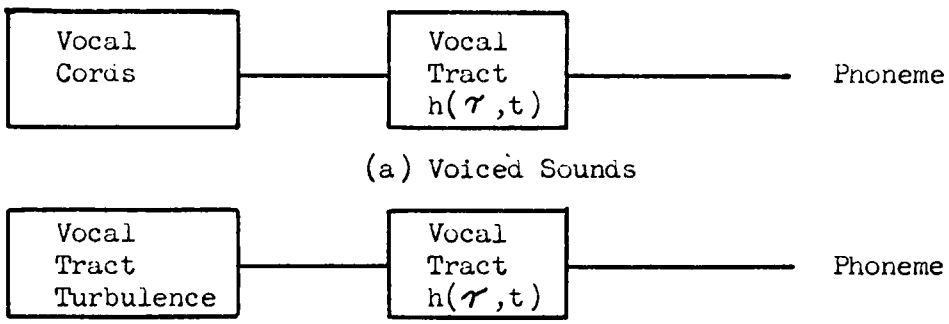
$$\int_{-\infty}^{\infty} |f(t)|\,dt$$

does not necessarily exist.[11] Hence, conventional Fourier analysis does not apply. A more useful description is found in the mathematically less rigorous concept of a time variant Fourier series.

A useful guide to the characteristics of speech might be found in the method of generation of speech. The voiced sounds, vowels and liquids, involve the vocal cords. The vocal cords vibrate, creating an essentially constant waveform, which is then modified by the

transfer function of the vocal tract. The unvoiced consonants are generated in a similar fashion, except that the sound generator is now turbulence in the vocal tract itself.[12] In the production of speech, modeled in Figure 2-2, both the generator and the vocal tract transfer function are functions of time.

Therefore, considering the frequency components of speech to be functions of time is intuitively reasonable. Alternatively, speech could be pictured as an infinite collection of sine wave oscillators whose amplitudes are functions of time.



(a) Voiced Sounds

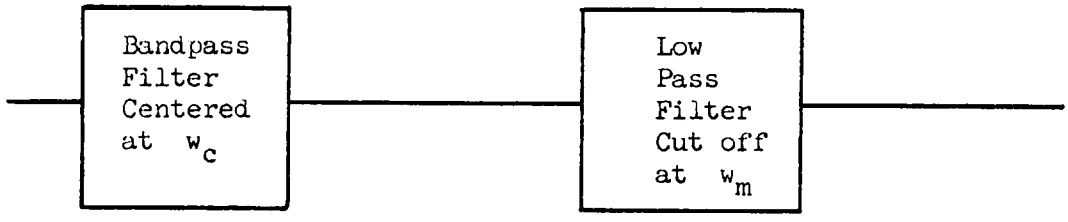
(b) Unvoiced Sounds

Model of Speech Generation

Figure 2-2

Now, rather than having a single phoneme described by a time invariant curve in the amplitude frequency plane, we picture continuous speech as a time variant curve in this plane. A particular phoneme is a particular pattern of variation of this curve. A vowel, for example, would be a curve with little variation over its time duration. The consonants, particularly the plosives, would vary quite rapidly.

The purpose of the speech processor is to specify this curve. This curve need not be specified completely, but only well enough to permit distinguishing one phoneme from another. A technique of accomplishing this would involve breaking the curve up into several intervals, and measuring the average amplitude over each interval. The circuit shown in Figure 2-3 provides the low frequency envelope of a band of frequencies, rather than a single frequency. The speech processor in Dr. Pickett's tactual vocoder is a number of these circuits connected in parallel, and provides a measure of the curve in the amplitude frequency plane.



Single Channel of Speech Processor

Figure 2-3

The output of the band pass filter in Figure 2-3 is speech, weighted by the impulse response of the band pass filter. Alternatively, if we view speech as the sum of infinite collection of sine wave oscillators whose amplitudes are functions of time, the output of the band pass filter is a weighted sum of this infinite collection. It is assumed that the band pass filters respond much more rapidly than the amplitude changes. Then, the rectifier and low pass filter provide the frequency envelope of this weighted sum.

The device shown in Figure 1-3 replaces this weighted sum by a single frequency. The low frequency output of each band pass filter--envelope detector is used to modulate a sinusoid whose frequency is the center frequency of the band pass filter. Then these modulated sinusoids are summed. Thus, the continuous time variant frequency spectrum of speech has been replaced by a finite number of frequencies. In effect, we have quantized the frequency spectra of speech.

As noted in Chapter I, the sum of these modulated sinusoids is synthesized speech. The intelligibility of this synthetic speech is a measure of how well this descrete spectra specifies speech.

The intelligibility of this synthetic speech is a function of the number of channels, the center frequencies of the band pass filters, and the frequency response of the band pass filters. For a given number of channels, the band pass filters could all have very wide band widths, or very narrow band widths, or a combination of wide and narrow band widths. Likewise, there are an infinite number of ways of assigning band pass filter center frequencies.

As it was not possible to vary all three of the above parameters, the number of channels was varied and two schemes were used for determining the other two parameters. It was noticed in some preliminary experiments that intelligibility seemed higher when the center frequencies were assigned logarithmically rather than linearly. So the center frequencies were always assigned in this manner, usually from 180 Hz to 4 KHz.

The band widths were adjusted according to two schemes. In one, the band widths were chosen so that the filter frequency characteristics overlapped at the half power points. In the other, the band widths were chosen as to fix the filter Q at 3. In all cases, the band pass filters were simple second order filters.

For a given configuration, random lists of vowels and consonants were read into the system and the output recorded. Then later the recording was listened to, and several subjects were tested. The percent correct recognition was taken as a measure of the adequacy of the approximation.

Chapter III

Filter Design

It was observed in Chapter I that the normal auditory system constitutes a pattern recognition device capable of providing an indication of the potential intelligibility of the speech processor output. As the measure of potential intelligibility used was the actual intelligibility of synthetic speech, this pattern recognition device needs little training. Each speech processor was evaluated by recording a list of synthesized phonemes, and later observing three subject's correct identification of these phonemes from this recording. By using this type of intelligibility testing, each speech processor need only be set up long enough for this list of phonemes to be recorded. Speech processor evaluation time was then dependent upon this recording time, and the time necessary to change the speech processor configuration. Thus, flexibility in the speech processor design was very desirable.

A single channel of the speech processor consists of a band pass filter, a detector, and a low pass filter. The characteristics of the low pass filters were fixed a priori, and only the band pass filter characteristic depended on the number of channels in the speech processor. Thus, flexibility in the band pass filters was of prime concern.

Passive L-C filters were ruled out as being too inflexible, and it was decided to use either analog computer implementation of a

second order filter or active R-C type filters. In either case, economy in number of amplifiers was necessary as in preliminary work speech processors with as many as sixteen channels were investigated. It was decided to use the facilities of the Sophomore Analog Computer Laboratory in the University of Arizona Department of Electrical Engineering. This laboratory consists of ten desk top $\pm$ 10 volt transistorized electronic differential analyzers, each with ten operational amplifiers. Of these ten amplifiers, four are committed as unity gain inverters, and the remaining six may be used as summers, integrators, or in an uncommitted feedback mode. Each computer has two quarter-square multipliers. When a multiplier is used, it requires committing two of the inverters and one of the uncommitted amplifiers. Thus, when both multipliers are in use, four uncommitted amplifiers remain on each computer. Each computer is equipped with five ten-turn and five single-turn potentiometers.

This facility combined the advantages of having a large number of available amplifiers and multipliers with occasional periods of time when no classes or other research activity utilized the computers. In the preliminary stages of this research over eighty of the hundred amplifiers were in use. Numerous circuit checkout problems made it necessary to have the speech processor set up for long periods of time.

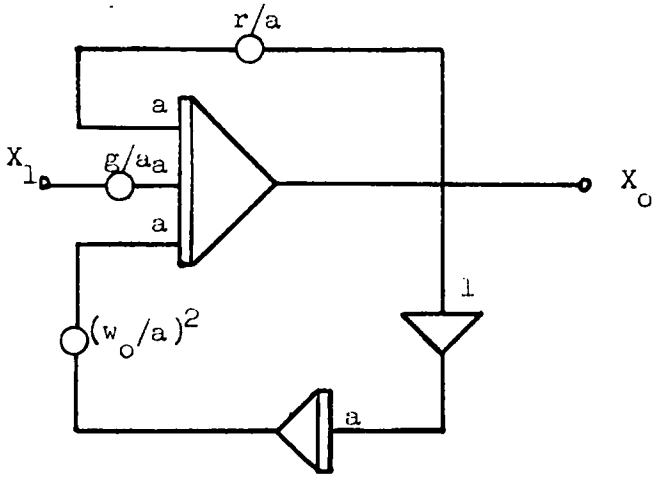
However, the Sophomore Analog Computer Laboratory has the disadvantage of not being primarily designed for research. The computer patch panels are designed with an eye to ease in patching elementary

problems rather than versatility. The computers are designed as slow differential analyzers. Furthermore, there was an appreciable and variable amount of high frequency noise on the computers. This noise was not detrimental as it was considerably above the audio range.

The first circuit to be considered for a second order band pass filter was the direct analog implementation of the transfer function

$$T(s) = \frac{-Ks}{s^2 + rs + w_o^2}$$

shown in Figure 3-1.[13] This circuit is extremely flexible, as the filter band width and center frequency can be varied directly and independently.



$$\frac{X_o}{X_1}(jw) = \frac{-jwg}{(w_o^2 - w^2) + jwr}$$

$$\text{Band Width} = \frac{w_o}{Q} \approx r$$

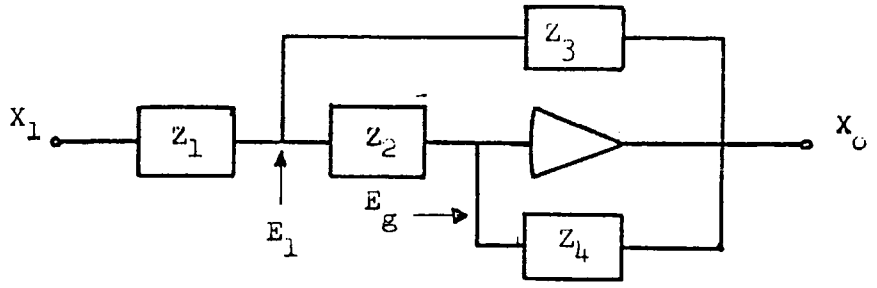
Analog Implementation of Second Order Filter

Figure 3-1

The integrators in the computers had ten μfd computing capacitors, which did not permit gains high enough to implement this circuit. For example, to implement a center frequency of 150 Hz requires input gain, a, greater than 1000. With ten μfd computing capacitors, this would require integrator input resistors less than 100 ohms. Furthermore, the current off set on the amplifiers was too high to permit use of smaller computing capacitors. This circuit also has the disadvantage of requiring two integrators and one inverter for each band pass filter. Thus for twelve channels, twenty four integrators and twelve inverters would be required. This circuit was removed from consideration since it required too many amplifiers and could not easily be implemented with the amplifiers available.

This leaves the possibility of realizing the second order filter by some type of active R-C circuit. Several practical considerations enter into the choice of this circuit. The frequency and band width of the realized transfer function should be conveniently and independently adjustable, preferably by varying a potentiometer. The circuit should have d.c. feedback to eliminate the drift problems encountered in the circuit of Figure 3-1. Finally, the number of amplifiers used should be kept to a minimum.

The second order filter transfer function, can be realized with one amplifier by a circuit of the form shown in Figure 3-2.[14]



Multiple Feedback Circuit

Figure 3-2

Writing the node equations (assuming amplifier gain infinite)

$$(X_1 - E_1)\frac{1}{Z_1} + (X_o - E_1)\frac{1}{Z_3} + (E_g - E_1)\frac{1}{Z_2} = 0$$

$$(E_1 - E_g)\frac{1}{Z_2} + (X_o - E_g)\frac{1}{Z_4} = 0$$

$$E_g = 0$$

which yields,

$$\frac{X_o}{X_1} = -\frac{Z_3 Z_4}{Z_2 (Z_1 + Z_2) + Z_1 (Z_3 + Z_4)} .$$

Considering Figure 3-2 further, it is seen that if either Z_1 or Z_2 is a pure capacitance, the transfer function must have the desired zero at the origin.

Since it is desired that the circuit have d.c. feedback, if Z_4 is a capacitance, Z_2 and Z_3 are resistances. If Z_2 or Z_3 are

capacitances, Z_4 must be a resistance. If Z_1 and Z_2 are capacitances, then Z_4 must be a resistance, to maintain d.c. feedback, and the transfer function would have an undesired pole at infinity. Thus, if Z_1 is a capacitance, Z_2 is a resistance. Alternatively, if Z_2 is a capacitance, Z_4 and Z_1 must be resistances.

Considering the first case, when $Z_1 = 1/sC_1$, $Z_2 = R_2$

$$\frac{X_o}{X_1}(s) = - \frac{Z_3 Z_4 s C_1}{Z_2 + s C_1 Z_3 R_2 + Z_3 + Z_4} .$$

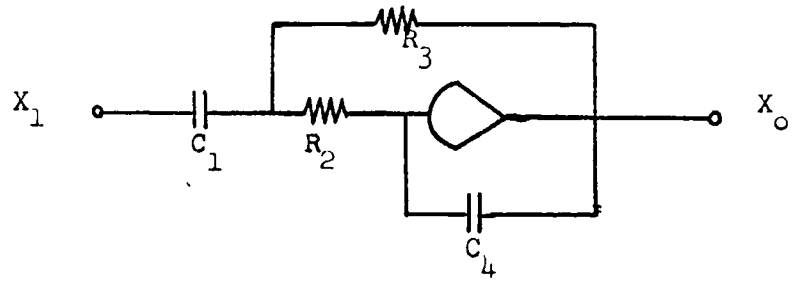
In order to maintain the desired d.c. feedback, both Z_3 and Z_4 can not be capacitances. If $Z_3 = 1/sC_3$, $Z_4 = R_4$,

$$- \frac{X_o}{X_1}(s) = \frac{s C_1 R_4}{s (C_3 R_2 + C_3 R_4 + C_1 R_2) + 1} ,$$

which is not the desired transfer function. However, if $Z_4 = 1/sC_4$, $Z_3 = R_3$

then,
$$- \frac{X_o}{X_1} = \frac{\frac{S}{R_2 C_4}}{s^2 + \frac{s (R_2 + R_3)}{C_1 R_2 R_3} + \frac{1}{C_1 C_4 R_2 R_3}}$$

thus, the active R-C circuit shown in Figure 3-1 uses one amplifier, has d.c. feedback, and realizes the desired transfer function.



$$\frac{X_o}{X_1} = - \frac{\dfrac{s}{R_2 C_4}}{s^2 + \dfrac{s\,(R_2 + R_3)}{C_1 R_2 R_3} + \dfrac{1}{C_1 C_4 R_2 R_3}}$$

Active R-C Filter

Figure 3-3

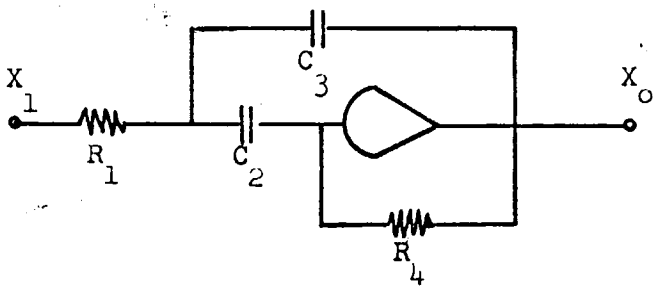
Alternatively, when $Z_2 = 1/sC_2$, $Z_1 = R_1$, $Z_4 = R_4$

$$\frac{X_o}{X_1} = - \frac{s\, C_2 Z_3 R_4}{R_1 + Z_3 + s\, C_2 R_1 (Z_3 + R_4)} .$$

If the desired transfer function is to be realized, Z_3 must be a capacitor $Z_3 = 1/sC_3$, and

$$\frac{X_o}{X_1} = - \frac{\dfrac{s}{C_3 R_1}}{s^2 + \dfrac{s\,(C_2 + C_3)}{R_4 C_2 C_3} + \dfrac{1}{C_2 C_3 R_1 R_4}} .$$

Thus, an alternate active R-C circuit using one amplifier, with d.c. feedback, realizing the desired transfer function is shown in Figure 3-4.[15]



$$\frac{X_o}{X_1} = - \frac{\dfrac{1}{C_2 C_3 R_1 R_4}}{s^2 + \dfrac{s\,(C_2 + C_3)}{R_4 C_2 C_3} + \dfrac{1}{C_2 C_3 R_1 R_4}}$$

Alternate Active R-C Filter

Figure 3-4

The circuit of Figure 3-3 was arrived at first, and was used for some preliminary trials. The transfer function for this circuit is

$$\frac{X_o}{X_1}(s) = - \frac{\dfrac{s}{R_2 C_4}}{s^2 + \dfrac{s\,(R_2 + R_3)}{R_2 R_3 C_1} + \dfrac{1}{C_1 C_4 R_2 R_3}}$$

If the center frequency is w_o, the band width B, and the midband gain G, then

$$w_o = (C_1 C_4 R_2 R_3)^{-1/2} \qquad \text{radians per second,}$$

$$B \approx \frac{R_2 + R_3}{C_1 R_2 R_3} \qquad \text{radians per second,}$$

and $$G = \left[\frac{C_1}{C_4}\right]\left[\frac{R_3}{R_2 + R_3}\right].$$

Adjusting either R_2 or R_3 changes both the band width and center frequency. The interaction was so severe that it was necessary to actually calculate values for C_1, C_4, R_2, R_3 for each filter, and make minor adjustments in R_2, R_3 to compensate for component tolerances. Typical values of C_1, C_4, R_2, R_3 for different filter center frequencies and band widths are shown in Table 3-1.

Table 3-1

Typical Component Values For Filter Of Figure 3-3

Center Freq. (Hz)	Band width (Hz)	C_1 (μfd)	C_4 (μfd)	R_2 (ohms)	R_3 (ohms)
250	120	0.1	.0047	25K	30K
400	220	0.1	.0047	10K	26K
650	280	0.1	.0047	10K	13K
800	400	0.1	.0047	10K	7.5K
1000	500	0.1	.0047	10K	4.5K
1250	500	0.1	.0047	10K	3.5K
1600	550	0.1	.001	10K	5.5K
2500	800	0.1	.001	10K	3.0K

The interaction encountered in the effects of R_2 and R_3 upon filter center frequency and band width made this circuit undesirable, and the circuit of Figure 3-4 was considered. The transfer function of this filter is

$$\frac{X_o}{X_1}(s) = - \frac{\dfrac{s}{C_3 R_1}}{s^2 + s\left(\dfrac{C_2 + C_3}{R_4 C_2 C_3}\right) + \dfrac{1}{C_1 C_2 R_1 R_4}}.$$

If the center frequency is w_o, and the band width B, then

$$w_o = (C_1 C_2 R_1 R_4)^{-1/2} \quad \text{radians per second, and}$$

$$B \approx \frac{(C_2 + C_3}{R_4 C_2 C_3}.$$

The center frequency depends upon C_1, C_2, R_1, R_4; but the filter band width is dependent only upon C_2, C_3, R_4. Therefore, adjustment of R_4 affects band width and center frequency, but adjustment of R_1 affects center frequency only. Furthermore, if Q is defined as the ratio of the filter center frequency to half power band width,

$$Q = \frac{w_o}{B},$$

then

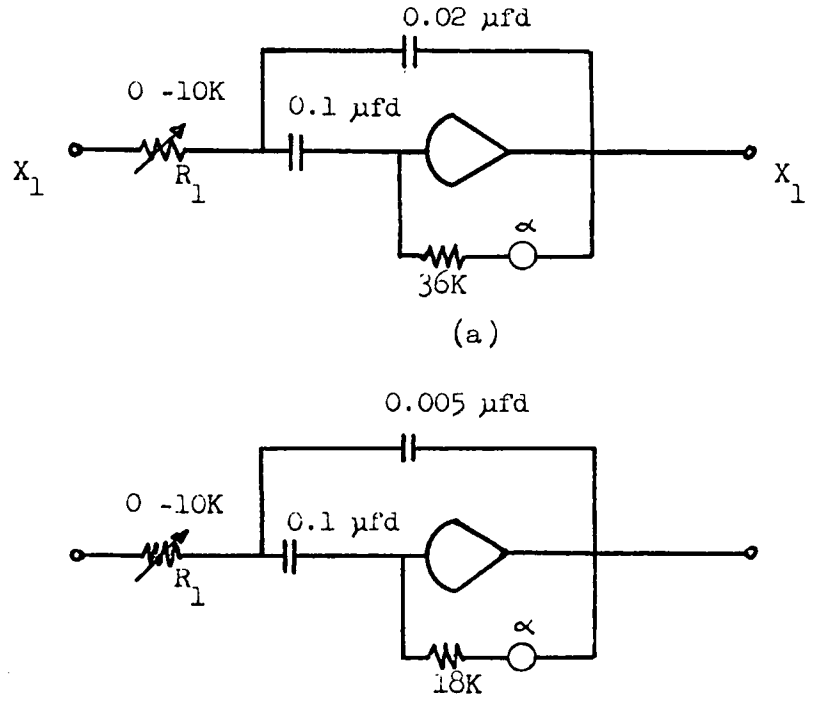
$$Q \approx \left[\frac{R_4}{R_1}\right]^{1/2} \times \frac{(C_2 C_3)^{1/2}}{C_2 + C_3}.$$

Since C_2 and C_3 are fixed, then for constant Q the ratio R_4/R_1 is fixed and the midband gain, G, equal to

$$\left[\frac{R_4}{R_1}\right]\left[\frac{C_2}{C_2 + C_3}\right]$$

is constant. This proved to be advantagous as the filter Q 's were usually nearly constant.

This particular implementation of the second order filter proved to be quite practical. Twelve of these filters were constructed in plug-in module form, using computer potentiometers as shown in Figure 3-5.



Band Pass Filters with Component Values

Figure 3-5

The circuit in Figure 3-5 (a) is designed to operate at filter center frequencies from 180 to 1000 Hz, with band widths appropriate to a filter Q from 1 to 3. The circuit in Figure 3-5 (b) covers the range from 1000 Hz to 5000 Hz with Q between 1 and 3. The transfer function for the circuit of Figure 3-5 (a) is

$$\frac{X_o}{X_1}(s) = - \frac{\left(\frac{5 \times 10^7}{R_1}\right) s}{s^2 + 1.66 \times 10^3 \alpha s + \frac{1.39 \alpha \times 10^{10}}{R_1}}$$

and for Figure 3-5 (b)

$$\frac{X_o}{X_1}(s) = - \frac{\left(\frac{2 \times 10^8}{R_1}\right) s}{s^2 + 1.17 \alpha \times 10^4 s + \frac{1.11 \times 10^{11} \alpha}{R_1}}$$

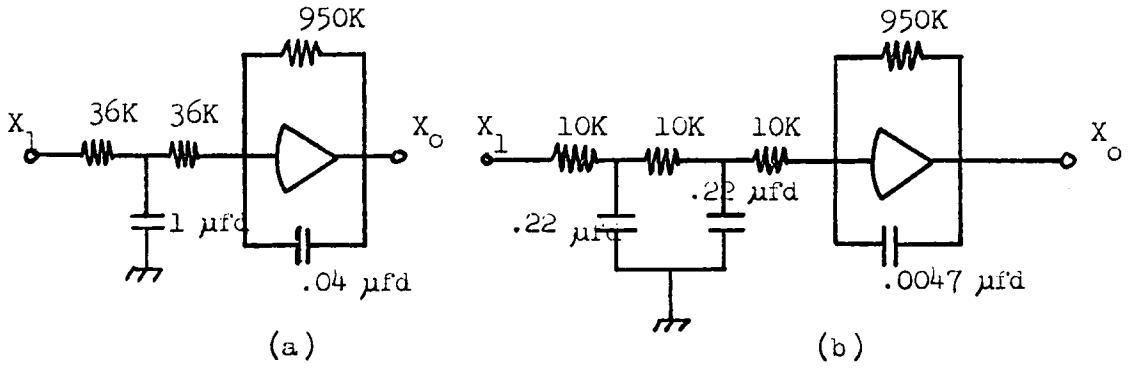
In practice, α and R_1 were set empirically, by simply setting α for the desired band width, and then adjusting R_1 for the desired center frequency. Note, however, that the midband gain is a function of both α and R_1. Thus, in general, the midband gain of the filters will not be identical. This will have to be compensated for elsewhere in the system.

It was stated earlier that the characteristics of the low pass filters were chosen from some a priori considerations. It is desired that the output of the lowpass filter be capable of varying as rapidly as the time variant spectrum. Controlled muscular movement is limited

to approximately ten Hz, so variations resulting from controlled changes in the vocal tract should be limited to ten Hz. However, faithful reproduction of the envelopes of the plosives is likely to require output frequencies higher than ten Hz. Further study of the effect of low pass filter cutoff frequencies on potential intelligibility is needed.

It was desired to fix the low pass filter cutoff frequency at some intuitively reasonable value. The low pass filter should be simple, use no more than one amplifier, and attenuate at least thirty db the center frequency of the band pass filter with which it is being used. Furthermore, it is desired that the gain of the low pass filters be adjustable in order to compensate for the differences in midband gain of the band pass filters. Then, the cutoff frequency should be as high as is commensurate with the preceding specifications.

The two types of low pass filters shown in Figure 3-6 were used.

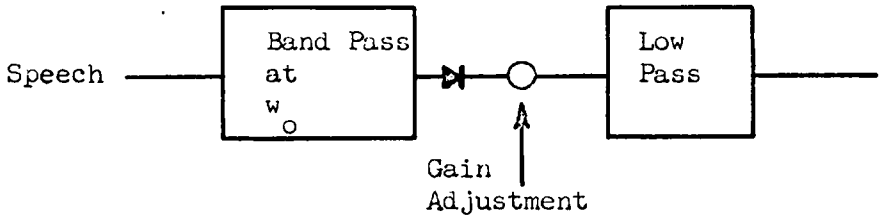


Low Pass Filters

Figure 3-6

The low pass filter shown in Figure 3-6 (a) is a two section type with a cutoff frequency of ten cps and twelve db/octave rolloff. The circuit in Figure 3-6 (b) had forty cps half power point with an eighteen db/octave rolloff. The low pass filter of Figure 3-6 (a) was used with the higher frequencies, while the circuit of 3-6 (b) was used at the lower frequencies.

A single channel of the speech processor appears in Figure 3-7.



Single Channel of the Speech Processor

Figure 3-7

The potentiometer was used to adjust the gain of the low pass filter to compensate for band pass filter gain differences.

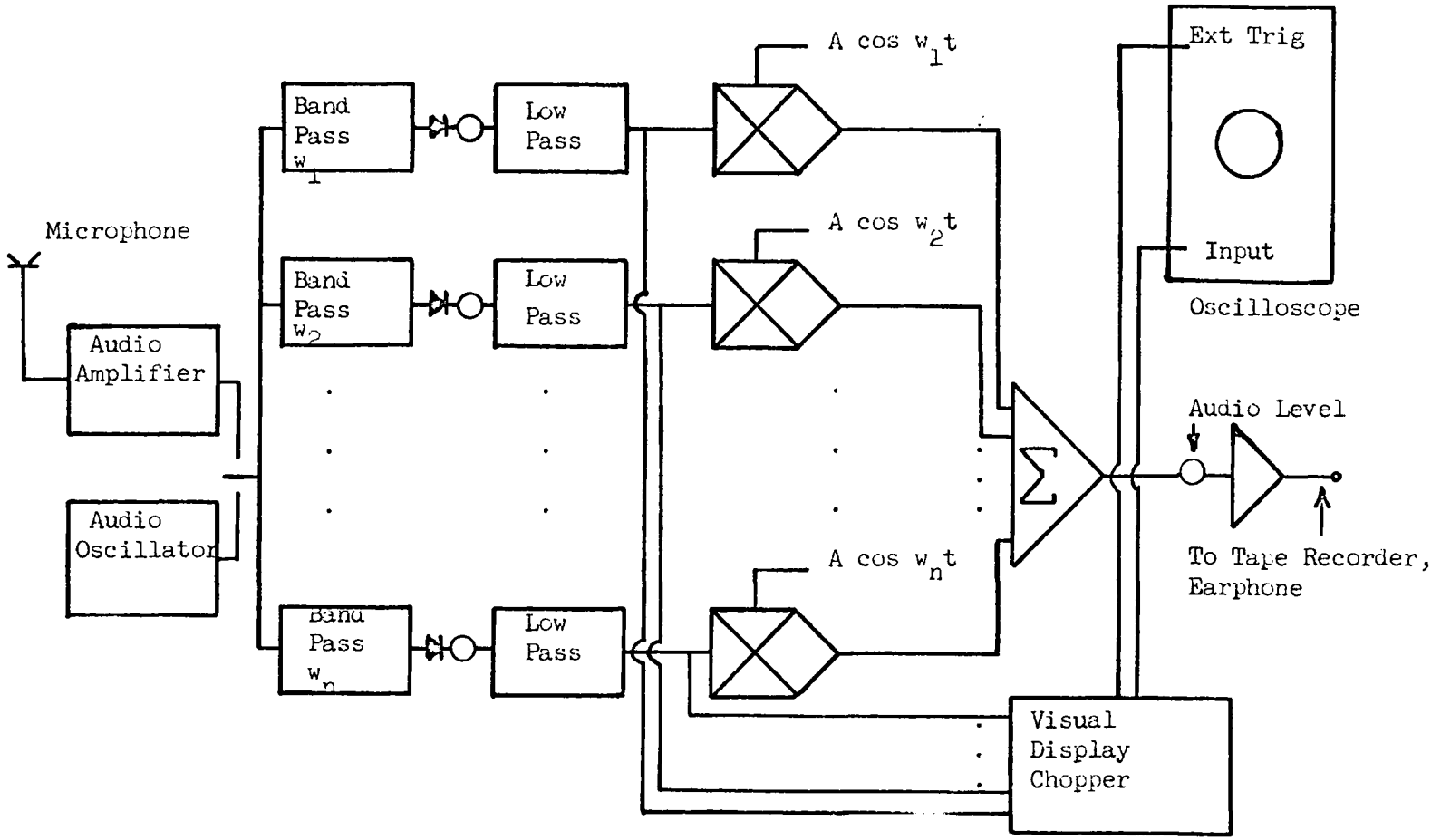
A single band pass filter-low pass filter combination requires two amplifiers. The multiplier used at the output required three amplifiers. Thus, a single channel of the speech processor-speech synthesizer required five amplifiers, and exactly two channels could be constructed on a ten amplifier computer.

Chapter IV

Test Procedure

In this study, the number of channels in the speech processor was varied from two to twelve. The output of this speech processor was used to synthesize speech using analog multipliers. The midband gain of each band pass filter used was a function of the center frequency and band width of the filter. Since both the center frequency and the band width of each filter was different, the midband gains of the filters used in any speech processor configuration were unequal. The center frequencies used were logarithmically equally spaced with band widths chosen to overlap approximately at the half power points. Fortunately, this method resulted in the midband gains being nearly equal, and it was possible to compensate for this variation in midband gains by making the d.c. gain of the low pass filters variable.

The actual adjustment of the speech processor channels for equal over all band pass-low pass gain was accomplished using the circuit of Figure 4-1 with the speech processor input connected to the audio oscillator. The visual display shown was designed and constructed by Mr. Donovan R. Biggs under the direction of Mr. Lorin P. Mc Rae as an undergraduate special project. This visual display sequentially presents up to sixteen voltages on a single oscilloscope sweep. Thus, it is possible to monitor continuously the outputs of all the low pass filters.



Complete Test Set-up with Visual Display

Figure 4-1

The oscillator amplitude was that necessary to drive the output of the band pass filter having the highest midband gain to the maximum allowable by the computer amplifiers, (twenty volts peak to peak). Then the oscillator frequency was varied to peak the output of the first low pass filter, and the output voltage of this low pass filter was adjusted to ten volts d.c.. This process was repeated for the remainder of the channels, assuring that the overall gains of the speech processor channels were equal.

Next, the speech synthesizer was adjusted. By changing the oscilloscope sweep speed, it was possible to monitor both the amplitude of the sinusoidal output of the multiplier and the d.c. output of the low pass filters. With the input audio oscillator amplitude at the same value as was used to adjust the speech processor, the oscillator frequency was varied to peak the output of the first low pass filter. Then the output of the first multiplier was adjusted to about one volt peak to peak. This process was repeated for the remainder of the multipliers. The speech synthesizer had to be adjusted by this procedure each time the speech processor configuration was changed since the audio oscillator output amplitude and the multiplier gain changed slightly with frequency.

The actual intelligibility testing was performed using the circuit of Figure 4-1 with the input connected to the audio amplifier. The speaker read text or a list of syllables into the microphone and the resultant synthesized speech was either recorded or monitored with

the earphones. Many early adjustments and modifications of the speech processor-speech synthesizer were made on subjective judgements of the intelligibility of the monitored output. The earphones used isolated the speaker and listener well enough that a given configuration performance could be informally evaluated quite simply. The speaker would read words or syllables into the microphone and the listener would respond verbally. In this manner, it was not only possible to quickly obtain a rough estimate of the intelligibility of the output, but to evaluate which particular phonemes were transmitted well, and which ones were not. It was by this type of testing that it was originally decided to arrange the center frequencies logarithmically equally spaced.

After it was ascertained that the system was performing properly, more objective testing was done. As outlined in Chapter I, the phoneme recognition rate was to serve as a measure of the potential intelligibility of the speech processor output. Thus, it was desired to measure the phoneme recognition rate with as little bias introduced by context as possible. Some of the phonemes are difficult to pronounce by themselves, so it was decided to form syllables with the phonemes. This was accomplished by separating the test into two parts. In the first, syllables were formed by pronouncing each consonant followed by the same vowel. The vowel used was "a" as in fat. A typical syllable would be ca, where the c is as the "c" in cat. These syllables were known as consonant syllables, and are listed in Table 4-1 (a).
In the second part, syllables were formed by pronouncing each vowel

Table 4-1

Vowel and Consonant Syllables

	(a)		(b)
1	sa	1	fee (even)
2	sha	2	fi (it)
3	cha	3	fa (fate)
4	va	4	fe (fed)
5	tha	5	fa (fat)
6	za	6	fa (all)
7	da	7	fo (folk)
8	ba	8	foo (foot)
9	ca	9	foo (food)
10	pa	10	fu (fuss)
11	ta		
12	ma		
13	la		
14	wa		
15	ra		
16	ja		
17	ga		
18	ya		
19	fa		
20	ha		

preceded by the same consonant (the fricative f). These syllables were known as vowel syllables and are listed in Table 4-1 (b). Random lists of twenty-five consonant syllables and fifteen vowel symbols were prepared and used for the testing.

Most of these syllables were not words in themselves and some occur rarely, if at all, as syllables in normal English language. This, and the knowledge that each consonant will be followed by the same, known vowel, and each vowel will be preceded by the same, known consonant effectively eliminates context. Thus each phoneme must be recognized by its own distinguishing features, unaided by contextual information. This does not eliminate the possibility that the recognizability of a phoneme might be a function of the phoneme immediately preceding or following it regardless of context. It was assumed that this effect, if it exists, would be small compared to the other parameters affecting intelligibility of the synthesized speech.

The lengths of these lists, twenty-five consonant syllables and fifteen vowel syllables, were necessarily a compromise. For statistical reasons, it was desired that the list be as long as possible. However, recording time and the length of time during which the subjects were available for testing was limited. For these reasons, the lists had to be kept reasonably short. It was decided to make the consonant syllable list long enough to contain each consonant at least once, and the vowel syllable list long enough to contain each vowel at least once. When the number of channels was reached where previous work had shown

the recognition rate to be changing rapidly with number of channels, several of these lists were recorded in order to improve the statistical reliability of the results. A typical consonant syllable list is shown in Table 4-2 (a) and a typical vowel syllable list in Table 4-2 (b).

Later, three subjects listened to these recorded lists and recorded their responses on a prepared test sheet. The test sheet is shown in Figure 4-2. The response to the vowel syllable was made by placing a mark in the appropriate row. The subject recorded their response to the spoken consonant symbol by writing their response next to the number of that consonant. These test sheets were scored by comparison with the lists spoken.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
fee (even)															
fi (it)															
fa (fate)															
fe (fed)															
fa (fat)															
fa (all)															
fo (folk)															
foo (foot)															
foo (food)															
fu (fuss)															

1 __ 11 __ 21 __

2 __ 12 __ 22 __

3 __ 13 __ 23 __

4 __ 14 __ 24 __

5 __ 15 __ 25 __

6 __ 16 __

7 __ 17 __

8 __ 18 __

9 __ 19 __

10 __ 20 __

Name _____

Test No.

V _____

C _____

Date _____

Vowels	Cons.

Test Sheet

Figure 4-2

Table 4-2

Typical Vowel and Consonant Syllable List

(a)

1	ha	14	ja
2	sa	15	na
3	ba	16	la
4	ca	17	ra
5	ma	18	wa
6	pa	19	ta
7	cha	20	da
8	va	21	tha
9	za	22	ha
10	ya	23	ja
11	sha	24	la
12	fa	25	ma
13	la		

(b)

1	fi (fit)
2	fa (fat)
3	fa (fate)
4	fe (fed)
5	fa (fall)
6	foo (food)
7	fu (fuss)
8	fee (even)
9	fa (fat)
10	fi (fit)
11	fa (all)
12	foo (foot)
13	fo (folk)
14	fu (fuss)
15	fa (fat)

Chapter V

Results

The intelligibility of the synthesized speech was determined by comparing the spoken phoneme lists with the subject's responses for each speech processor configuration. In each speech processor, the band pass filter center frequencies were assigned in a logarithmically equally spaced manner. For speech processors with twelve channels through five channels, the filter band widths were chosen to overlap at approximately the three db points. The filter center frequencies and band widths for these are listed in Tables 5-1 through 5-8. Then, for speech processors with eight channels through three channels, the center frequencies were chosen in the same manner and the Q of the band pass filters was held constant at three, as shown in Tables 5-9 through 5-14.

The assignment of the band pass filter center frequencies in a logarithmically equally spaced manner, with overlapping half power points, resulted in nearly constant Q filters, as can be observed from Tables 5-1 through 5-9. Assignment of the band pass filter characteristics in this manner resulted in increasingly broad filters as fewer channels are used. Thus a large segment of the curve in the frequency-amplitude plane falls into the pass band of a single filter. When the Q of the band pass filters was held constant at three, better frequency

resolution was obtained when less than ten channels were used. However, this increase in frequency resolution resulted in segments outside the pass band of any filter.

For example, if the constant Q case and the overlapping half power point case are compared for a five channel speech processor, the distinction becomes clear. Table 5-8 shows the filter characteristics for the overlapping half power point configuration, and Table 5-12 shows the filter characteristics for the constant Q configuration. The constant Q filters are narrow and widely spaced compared to the overlapping half power point filters.

While constant Q filters with more than eight channels were not investigated, it can be seen from Table 5-3 that the two cases are nearly the same for ten channels. For more than ten channels, the constant Q filters would yield poorer frequency resolution than the overlapping half power point filters.

Results for the overlapping half power point filters are tabulated in Table 5-15, and for the constant Q filters in Table 5-16. Figure 5-1 shows the intelligibility of vowels and consonants versus number of speech processor channels for the constant Q case. Figure 5-2 shows the corresponding plot for the overlapping half power point case. Figure 5-3 shows the overall phoneme intelligibility for both filter configurations versus number of channels.

Figure 5-3 shows that the greater frequency resolution afforded by the filters with Q of three resulted in somewhat higher intelligibility.

If the curve for the constant Q case is extrapolated to nine or ten channels it appears that the two curves approach each other. This was to be expected since the two cases are nearly the same for ten channels.

Figure 5-3 shows that the intelligibility at seven channels for the overlapping half power point filters is noticeably less than might be expected from the other points. Referring to Figure 5-2, it is seen that vowel intelligibility was unexpectedly low for this case. The reason for this is not known. Reference to the raw data (not included) shows that the scores of all three subjects on vowel recognition were nearly the same for this case. Hence, this low value being due to statistical fluctuation seems unlikely. It might have been caused by some undetected malfunction in the test set-up.

From Figures 5-1 and 5-2, it is noted that vowel intelligibility for greater than five channels is greater than consonant intelligibility. However, Figure 5-1 indicates that vowel intelligibility for less than five channels is less than consonant intelligibility. This might be taken to indicate that the intelligibility of consonants, to a greater extent than the intelligibility of vowels, is dependent on time variance rather than frequency content.

Table 5-1

Band Pass Filters for Twelve Channel Speech Processor

Center Freq. (Hz)	Band Width (Hz)	Q
180	60	3
240	70	3.4
320	90	3.5
420	120	3.5
560	160	3.5
750	220	3.4
1000	275	3.6
1300	375	3.5
1750	500	3.5
2300	675	3.4
3100	950	3.3
4200	1100	3.8

Table 5-2

Band Pass Filters for Eleven Channel Speech Processor

Center Freq. (Hz)	Band Width (Hz)	Q
180	70	2.6
250	80	3.1
335	100	3.4
455	150	3.0
625	200	3.1
855	300	2.8
1200	375	3.2
1600	500	3.2
2200	700	3.1
3000	1000	3.0
4200	1200	3.5

Table 5-3

Band Pass Filters for Ten Channel Speech Processor

Center Freq. (Hz)	Band Width (Hz)	Q
180	75	2.4
255	90	2.8
360	125	2.9
510	180	2.8
730	250	2.9
1050	375	2.8
1500	500	3.0
2100	750	2.8
3000	1050	2.8
4200	1200	3.5

Table 5-4

Band Pass Filters for Nine Channel Speech Processor

Center Freq. (Hz)	Band Width (Hz)	Q
180	85	2.1
265	100	2.6
390	155	2.5
580	230	2.5
850	330	2.6
1250	450	2.8
1750	700	2.5
2700	1200	2.2
4200	1500	2.8

Table 5-5

Band Pass Filters for Eight Channel Speech Processor

Center Freq. (Hz)	Band Width (Hz)	Q
180	100	1.8
280	130	2.1
435	200	2.1
670	300	2.2
1050	500	2.1
1620	800	2.0
2600	1300	2.0
4200	1600	2.5

Table 5-6

Band Pass Filters for Seven Channel Speech Processor

Center Freq. (Hz)	Band Width (Hz)	Q
180	120	1.5
300	160	1.9
510	280	1.8
860	470	1.8
1450	820	1.8
2500	1400	1.8
4200	1700	2.5

Table 5-7

Band Pass Filters for Six Channel Speech Processor

Center Freq. (Hz)	Band Width (Hz)	Q
180	120	1.5
340	210	1.6
610	430	1.4
1200	750	1.6
2100	1500	1.4
4200	2100	2.0

Table 5-8

Band Pass Filters for Five Channel Speech Processor

Center Freq. (Hz)	Band Width (Hz)	Q
180	110	1.6
400	340	1.2
870	540	1.6
1950	1600	1.2
4200	2250	1.9

Table 5-9

Band Pass Filters for Eight Channel Speech Processor with Q = 3

Center Freq. (Hz)	Band Width (Hz)
180	60
280	90
435	145
670	220
1050	350
1620	540
2600	900
4200	1500

Table 5-10

Band Pass Filters for Seven Channel Speech Processor with Q = 3

Center Freq. (Hz)	Band Width (Hz)
180	60
300	100
510	160
860	290
1450	500
2500	830
4200	1500

Table 5-11

Band Pass Filters for Six Channel Speech Processor with Q = 3

Center Freq. (Hz)	Band Width (Hz)
180	60
340	110
610	200
1200	400
2100	700
4200	1500

Table 5-12

Band Pass Filters for Five Channel Speech Processor with Q = 3

Center Freq. (Hz)	Band Width (Hz)
180	60
400	130
870	300
1950	650
4200	1500

Table 5-13

Band Pass Filters for Four Channel Speech Processor with Q = 3

Center Freq. (Hz)	Band Width (Hz)
180	60
510	170
1450	500
4200	1500

Table 5-14

Band Pass Filters for Three Channel Speech Processor with Q = 3

Center Freq. (Hz)	Band Width (Hz)
180	60
900	300
4200	1500

Table 5-15

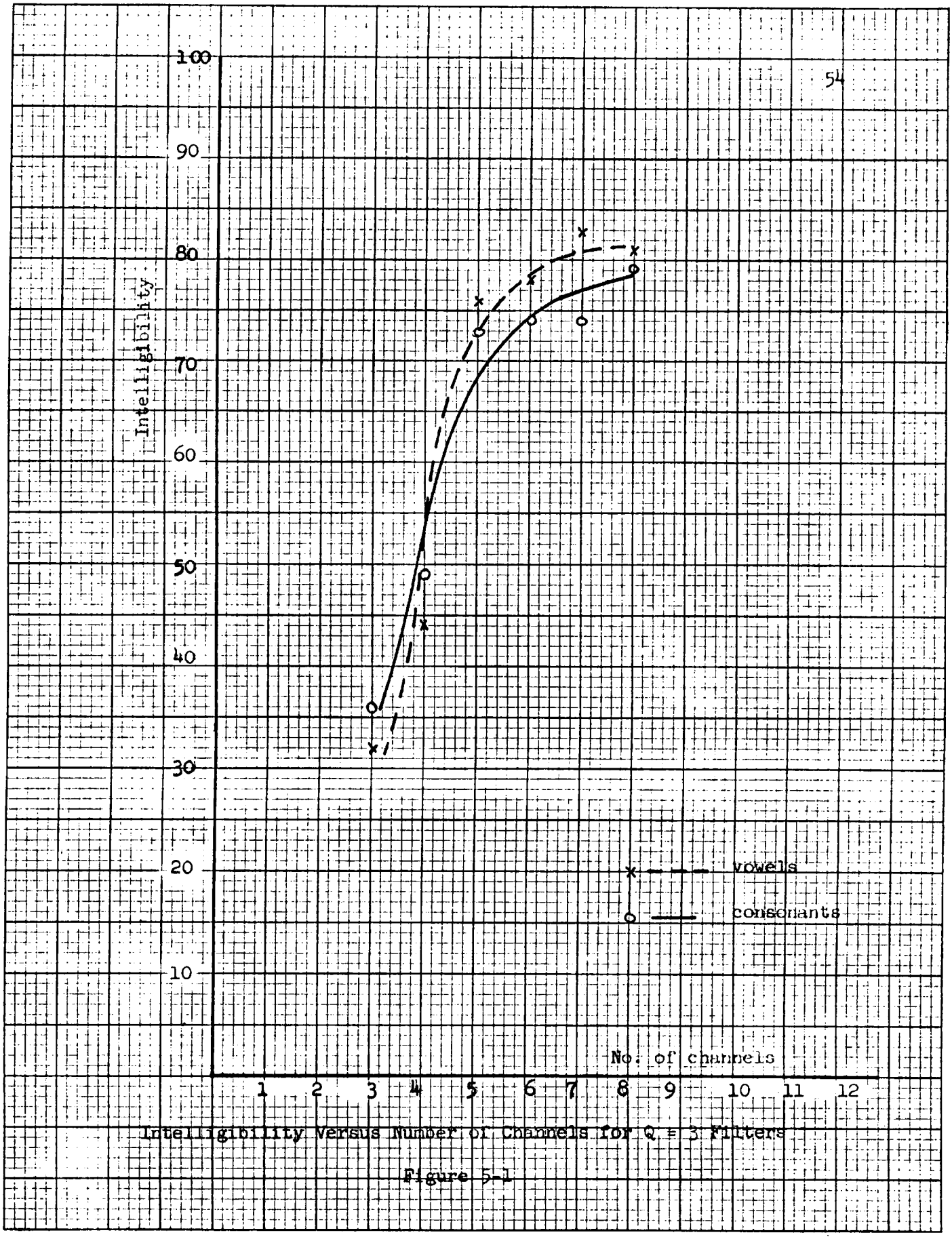
Intelligibility for Speech Processors With Overlapping Half Power Point Filters

	Vowels		Consonants		Overall	
No. of Channels	No. Tested	Per Cent Correct	No. Tested	Per Cent Correct	No. Tested	Per Cent Correct
12	270	85	450	78	720	81
11	180	85	300	77	480	80
10	180	88	300	74	480	79
9	180	84	300	74	480	78
8	180	83	300	73	480	77
7	180	64	300	72	480	69
6	90	70	150	69	240	70
5	90	68	150	55	240	60

Table 5-16

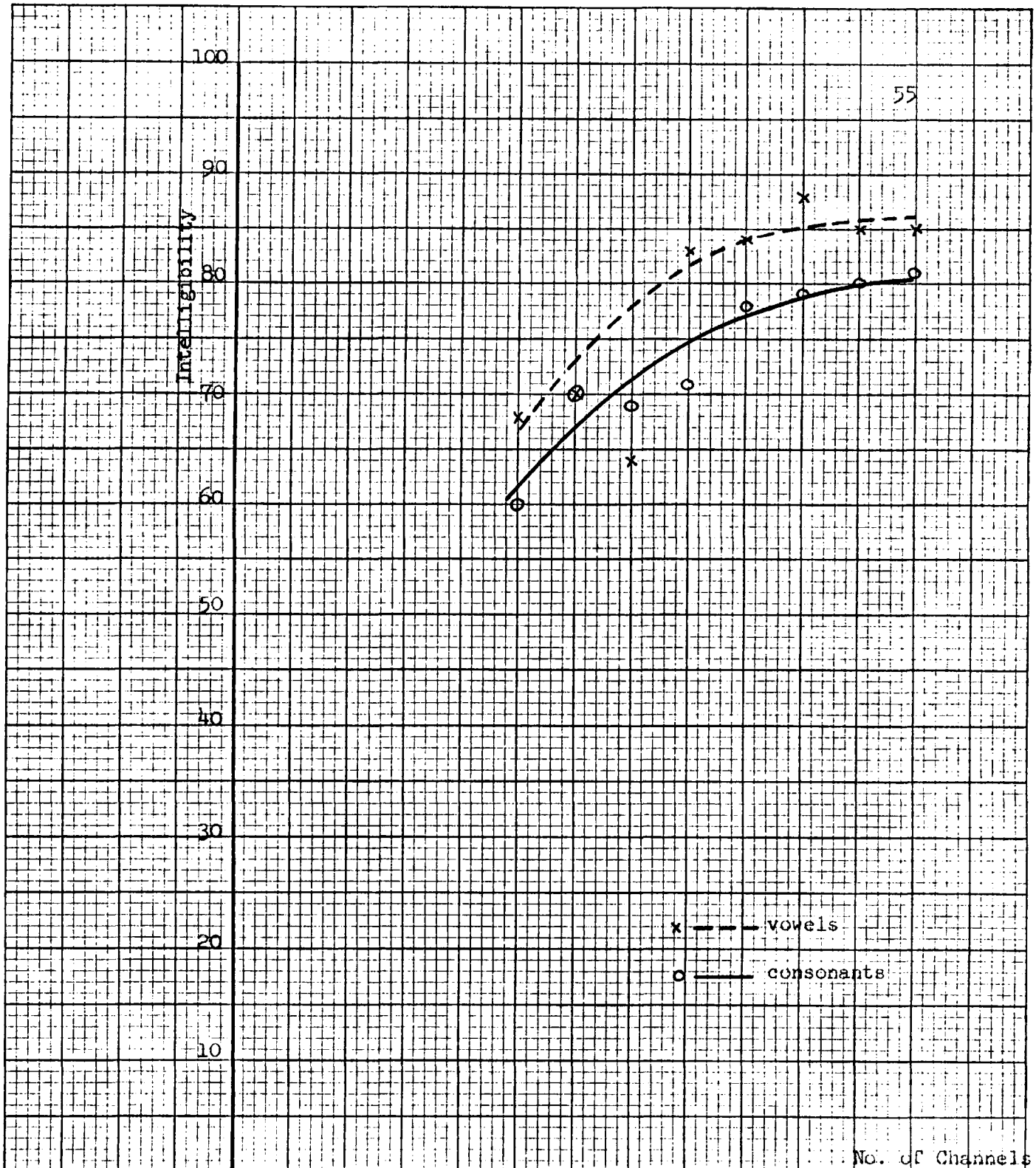
Intelligibility for Speech Processors With Band Pass Filters with Q = 3

	Vowels		Consonants		Overall	
No. of Channels	No. Tested	Per Cent Correct	No. Tested	Per Cent Correct	No. Tested	Per Cent Correct
8	315	81	400	77	715	79
7	240	83	400	69	640	74
6	240	78	400	71	640	74
5	240	76	400	72	640	73
4	240	44	400	53	640	49
3	90	32	150	38	240	36



Intelligibility Versus Number of Channels for Q = 3 Filters

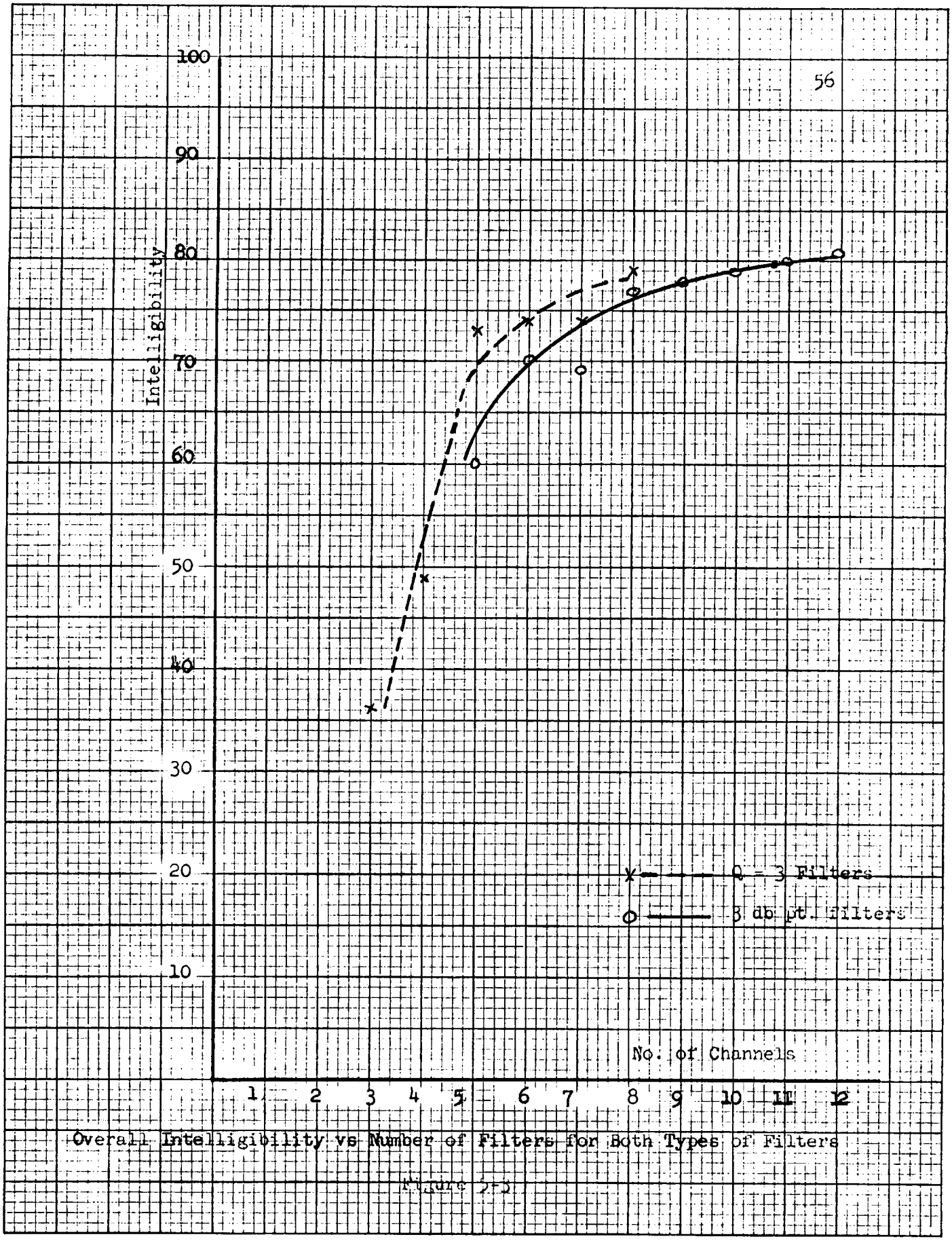
Figure 5-1



Intelligibility Versus Number of Channels

For Overlapping 3 db. Point Filters

Figure 5-2



Overall Intelligibility vs Number of Filters for Both Types of Filters

Figure 5-3

Chapter VI

Further Topics

In Chapter II, a technique of approximating the continuous time variant frequency spectrum of speech by a finite number of frequencies was discussed. This approximation allowed the amplitude of each of these frequencies to vary continuously. A further simplified description could be obtained if the amplitudes of these frequencies were restricted to descrete levels. Than the curve in the frequency-amplitude plane would be restricted to a finite number of configurations. For example, if each frequency was allowed to assume m amplitudes, and a n channel speech processor were used, then there would be m^n possible distinguishable curves. However, since time is still a continucus parameter, an infinite number of time variant patterns still exist.

It is likely that the skin can detect only a finite number of levels of stimulation. Thus, in effect, the skin may act as a quantizer on the output of the speech processor. The number of levels of stimulus levels distinguishable by the skin is not known at this time. However, since the skin may act as a quantizer, the effect of quantization on the potential intelligibility of the speech processor output is of interest.

It is possible that the skin is capable of detecting only on-off type stimuli. If quantized speech processor outputs were available, a separate stimulator could be assigned to each level of quantization.

Thus, the amplitude of a given channel would be represented to the tactile nervous system by the position of the stimulus, rather than by the amplitude of the stimulus. This stimulation scheme would emphasize the problem of minimizing the number of channels and number of levels of quantization used, due to the power, size, and weight requirements of the larger number of stimulators.

Assuming that the output quantization may be of value, the philosophy of speech processor design outlined in Chapter I applies, with a few additional parameters in question. That is, the potential intelligibility of the speech processor output must be adequate if satisfactory operation is to be realized. In addition to the number of speech processor channels, and the frequency allocation of these channels, the quantization scheme now affects the potential intelligibility of the output. It is not immediately clear whether each channel should be assigned the same number of quantization levels, or how the output of a given channel should be quantized.

Some preliminary work was done with output quantization of a eight channel speech processor. The band pass filters used had center frequencies assigned logarithmically equally spaced from 180 Hz to 4 KHz, with Q = 3, as shown in Figure 5-9. A number of quantizers were designed and built by Mr. Robert Darwin under the direction of Dr. Fredrick Hill as an undergraduate special project. These quantizers allowed the output of a speech processor channel to assume levels of approximately, 0, 2.5, 5, 7.5 or 10 volts. When these quantizers were

inserted directly between the low pass filter outputs and the analog multipliers, an overall recognition rate of 55.5% was observed. The vowel recognition rate was 76.5% and the consonant recognition rate was 43.5%.

It was noticed by observing the visual display that the quantizer outputs in the lower frequency ranges were assuming effectively two states. Either the amplitude at these frequencies was not high enough to reach the threshold of the first level, or it was high enough to achieve the highest level. A possible explanation for this behavior is to be found in the amplitude characteristics of speech. The vowels and voiced consonants have large components in these frequency ranges, while the unvoiced sounds tend to have lower levels concentrated in the higher ranges.

Ordinarily, the d.c. output of a given low pass filter, when a sinusoidal input at the center frequency of the corresponding band pass filter is applied, is equal to the d.c. output of any second low pass filter, when a sinusoidal input of the same amplitude at the center frequency of the second band pass filter is applied. In an attempt to gain more effective quantization levels at the lower frequencies, the outputs of the low pass filters were weighted prior to quantization. When a constant amplitude sinusoid was applied to the input of the speech processor, the gain of the low pass filters was adjusted to produce the output levels shown in Figure 6-1. For example, if the output of the low pass filter corresponding to the 2600 Hz band pass

filter were eight volts for an input of 2600 Hz, the output of the low pass filter corresponding to the 435 Hz band pass filter to a sinusoidal input of the same amplitude and a frequency of 435 Hz would be four volts. With this scheme, the vowel recognition rate rose to 90% and the consonant recognition rate to 66%. The overall recognition rate was 75%.

Band Pass Filter Center Freq. (Hz)	Low Pass Filter Output
180	6
280	5
435	4
670	6
1050	8
1620	8
2600	8
4200	8

Low Pass Filter Weighting

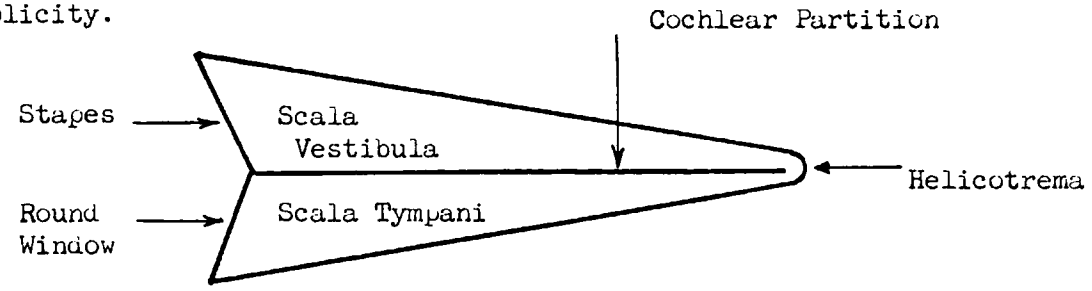
Figure 6-1

The preceding results should not be considered quantitatively accurate, as the number of phonemes tested was small. In the first case, three subjects listened to two lists twice each. In the second

case, three subjects listened to one list twice each. However, these results do indicate that the different channels should not necessarily be dealt with uniformly. Exactly how the outputs should be quantized will be the object of further research, with the answer perhaps to be found in the probability distributions of different speech frequencies.

The design of the speech processor used in these experiments arose from consideration of the characteristics and production of speech. An alternative speech processor results from consideration of the actual hearing apparatus.

A simplified diagram of the cochlea appears in Figure 6-2. The cochlea is normally coiled like a snail, but is shown uncoiled for simplicity.



Simplified Diagram of the Cochlea

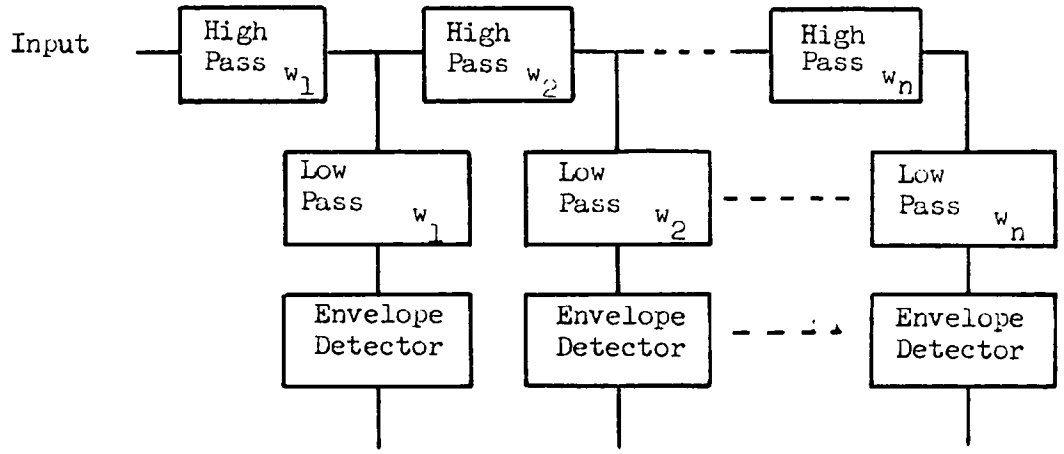
Figure 6-2

The cochlea is filled with a colorless liquid, perilymph, which transmits the vibrations introduced by the stapes. The cochlear partition is a channel, filled with a liquid called endolymph, bounded by the basilar membrane, Reissner's membrane, and a bony shell. Audio frequency vibrations are transmitted from the scala vestibula to the scala

tympani through the cochlear partition. This vibration is introduced into the auditory nerve by the organ of Corti, which contains the auditory nerve endings.

The point on the basilar membrane where the greatest amplitude vibration occurs depends upon the stimulus frequency. As the stimulus frequency increases, this point of greatest amplitude vibration of the basilar membrane moves away from the stapes toward the helicotrema. The frequency response of a single point on the basilar membrane approximates that of a band pass filter.[16]

A speech processor designed with the operation of the cochlea in mind appears in Figure 6-3.



Serial Filter Speech Processor

Figure 6-3

This speech processor simulated the serial filter nature of the cochlea. A twelve section version of this speech processor was designed and built by Mr. Raymond Wilsey as an undergraduate special project under the direction of Mr. Lorin Mc Rae. A preliminary test using six sections of the serial filter was performed. Using the envelope detectors and the speech synthesis method which was used with the parallel filter speech processor, vowel intelligibility was 76% and consonant intelligibility was 61%. The intelligibility of the vowels and consonants combined was 66%. The break frequencies w_1 through w_6 were assigned logarithmically equally spaced from 200 Hz to 4000 Hz.

The above results indicate that the series filter arrangement provided about the same output intelligibility as the parallel filter speech processor. Furthermore, since this speech processor is intended to replace the cochlea, it may be advantageous for it to simulate the operation of the cochlea. What the advantages are, if any, of the serial filter speech processor over the parallel filter speech processor is not known at this time.

In this study, intelligibility was defined as the percentage correct identification of phonemes. This definition was adopted to eliminate, as much as possible, contextual influence of sound recognition. However, this restriction resulted in the testing of a limited, known vocabulary. In order to obtain an indication of the effects of increasing this vocabulary, with an increase in context, a list of one hundred words was prepared. This list, shown in Table 6-1, was read

into an eight channel, Q = 3, speech processor, and the synthetic speech output recorded. The percentage correct recognition of this enlarged vocabulary was 58%. Three subjects listened to this recording twice each.

As indicated in Chapter I, word recognition of 60% corresponds to sentence recognition of over 95%. Thus, while word recognition found in this test was not as good as phoneme recognition for the same speech processor, it is probably adequate for sentence recognition.

Table 6-1

Word List

1	thin	21	azure	41	even	61	use	81	he
2	let	22	drink	42	bed	62	care	82	cut
3	top	23	chin	43	had	63	ten	83	did
4	old	24	fat	44	ton	64	father	84	get
5	turn	25	date	45	put	65	dear	85	tool
6	bed	26	did	46	vat	66	car	86	focus
7	fall	27	fall	47	arch	67	over	87	sanity
8	yard	28	trim	48	then	68	put	88	note
9	zebra	29	will	49	troth	69	trim	89	this
10	father	30	sell	50	ring	70	kill	90	can
11	ape	31	date	51	arch	71	bite	91	best
12	go	32	under	52	yard	72	mile	92	as
13	had	33	fur	53	drink	73	bake	93	between
14	not	34	off	54	haze	74	tone	94	sound
15	she	35	get	55	have	75	horn	95	and
16	always	36	leisure	56	hat	76	joy	96	lips
17	meet	37	dash	57	pass	77	ahead	97	trying
18	bite	38	lap	58	red	78	fork	98	sound
19	pass	39	bear	59	tool	79	out	99	from
20	yet	40	car	60	doubt	80	troop	100	herd

Chapter VII

Conslusions

The object of the Artificial Ear project is to enable the deaf to recognize speech. As outlined in Chapter I, this is to be accomplished by replacing the cochlea by a speech processor, and by substituting the tactile nerves for the auditory nerve. A primary requirement for the success of such a scheme is that the potential intelligibility of the speech processor output be sufficient for reasonable speech recognition.

Speech processor output intelligibility was determined in this study by the intelligibility of synthetic speech. In this manner, the normal hearing apparatus was used as a pattern recognition device. This pattern recognition device is not ideal, and the intelligibility of this synthetic speech is not potential intelligibility as defined in Chapter I. However, this pattern recognition device is as near ideal as is presently available. Thus, for practical purposes, the intelligibility of this synthetic speech is an upper bound on the recognition of speech from the speech processor outputs. At best, tactile recognition of speech from the speech processor outputs will approach the recognition rate of this synthetic speech.

Figure 5-3 indicates that intelligibility does not rapidly increase when more than six to eight speech processor channels are used.

Word recognition for eight channels was 58%, which is adequate for about 90% sentence recognition.

In this study, the method of arranging the filter center frequencies and band widths was chosen from the informal results of preliminary testing. Some other method for arranging the speech processor channels may yield even better output intelligibility for a given number of channels than was found here. Such a method might arise from further considerations of either speech and its production, or hearing. Further study on the effects of low pass filter frequency characteristic upon output intelligibility might lead to improved output intelligibility.

Further time could be spent on improving the accuracy of these results. Larger phoneme lists could be used, and more subjects tested. However, this would probably not be in the best interests of the Artificial Ear project. This study has indicated that six to eight channels are adequate, and further work should be directed towards optimizing a speech processor with this number of channels.

One approach to optimizing a speech processor with a given number of channels involves the amplitude probability distributions of speech at different frequencies. If these amplitude distributions could be determined, it may be possible to arrange the filter band widths and center frequencies so that the average information transmitted through any channel would be equal to the average information transmitted through any other channel. In this manner, each stimulator would be, on the average, contributing an equal amount to the tactile

recognition of speech.

At this point, perhaps the major effort should be extended towards determination of the stimulation characteristics of the skin. The preliminary results of the study on speech processor output quantization indicate that proper output quantization should not result in a drastic reduction in potential intelligibility. Hence, the extent of quantization imposed by the skin should be determined, and the speech processor output quantized accordingly.

For example, if it should result that it is only possible to tactually detect three distinct levels of stimulation, then the speech processor output should be quantized so as to maximize the average information contained in each level. Maximization of the average information contained in each level will probably result from the previously mentioned amplitude probability distributions.

Continuing research on tne Artificial Ear project should proceed along two lines. The amplitude probability distributions at different frequencies for speech should be determined, with the idea in mind of optimizing the output potential intelligibility of a six to eight channel speech processor with possibly quantized outputs. At the same time, effort should be extended toward determination of the stimulation characteristics of the skin, and the design and construction of adequate tactile stimulators.

LIST OF REFERENCES

1. Dudley, Homer "The Vocoder," Bell Labs Record, Vol. 18, 1939, pp. 122-126.

2. Flanagan, J.L. Speech Analysis, Synthesis and Perception, Academic Press Inc., New York, 1965, p. 248.

3. Barney, H.L., Haworth, F. E., and Dunn, H.K. "An Experimental Transistorized Artificial Larynx", BTL Monograph No. 3395 p.19.

4. Pickett, J.M. "Tactual Communications of Speech Sounds to the Deaf: Comparison with Lip Reading," Journal of Speech and Hearing Disorders, Vol. 28, No. 4, Nov. 1963, p. 315.

5. Pickett, J.M. and Pickett, B.H. "Communication of Speech Sounds by a Tactual Vocoder," Journal of Speech and Hearing Research, Vol. 6, No. 3, September 1963, p. 207.

6. Flanagan, op. cit. p. 247.

7. Dudley, Ibid.

8. Lepage, W.R. Complex Variables and the LaPlace Transform for Engineers, Mc Graw-Hill, New York, 1961, p. 268.

9. Fletcher, H.D. Speech and Hearing in Communication, D. Van Nostrand, New York, 1953, p. 288.

10. Manley, H.J. "Analysis-Synthesis of Connected Speech in Terms of Orthogonalized Exponentially Damped Sinusoids," Journal of Acoustical Society of America, Vol. 35, No. 4 April 1963, p.473.

11. Flanagan, op. cit., p. 120.

12. Flanagan, op. cit., pp. 9-21.

13. Korn, G.A. Random Process Simulation and Measurements, Mc Graw-Hill, New York, 1966, p. 197.

14. Jeness, R. R., Analog Computation and Simulation: Laboratory Approach, Allyn and Bacon, Boston, 1965, p. 87.

15. Korn, G.A., Ibid.

16. Flanagan, op. cit., pp. 81-85.