Computerized speech analysis

by

Mark Byron Wehde

A Thesis Submitted to the

Graduate Faculty in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major: Biomedical Engineering

Iowa State University
Ames, Iowa

1984

TABLE OF CONTENTS

INTRODUCTION

Data Reduction

One problem in speech recognition is extracting significant features from the speech waveform. If such a waveform can be simplified without reducing its informational content, the feature extraction process should be simpler. Early experiments by Licklider and Pollack (1948) have shown that significant data reduction can be achieved by dichotomization without a concomitant loss in intelligibility. Because the resulting waveform is a two-valued function of time, it lends itself to modern digital analysis techniques, either through special purpose hardware or mini and microcomputer applications.

The frequency spectrum of the speech waveform is thought to contain distinctive features. This is because the vocal tract is a resonant cavity with resonant properties which vary according to the configuration of the articulatory organs; referring to the tongue, lips, jaw, and velum; according to Holmes (1972). Figure 1C shows a typical acoustical waveform. The resonant frequencies are labeled. The resonant peaks caused by concentrations of acoustic energy at certain frequencies are known as formants (Flanagan, 1965). Many researchers feel that the ranges of the first two formants of speech contain the information necessary for recognition of many speech sounds (De Mori, 1971; Thomas, 1968). In fact, many speech sounds can be identified on the basis of the locations of their first and second formants. These are the features which researchers have tried to isolate via zero-crossing analysis.

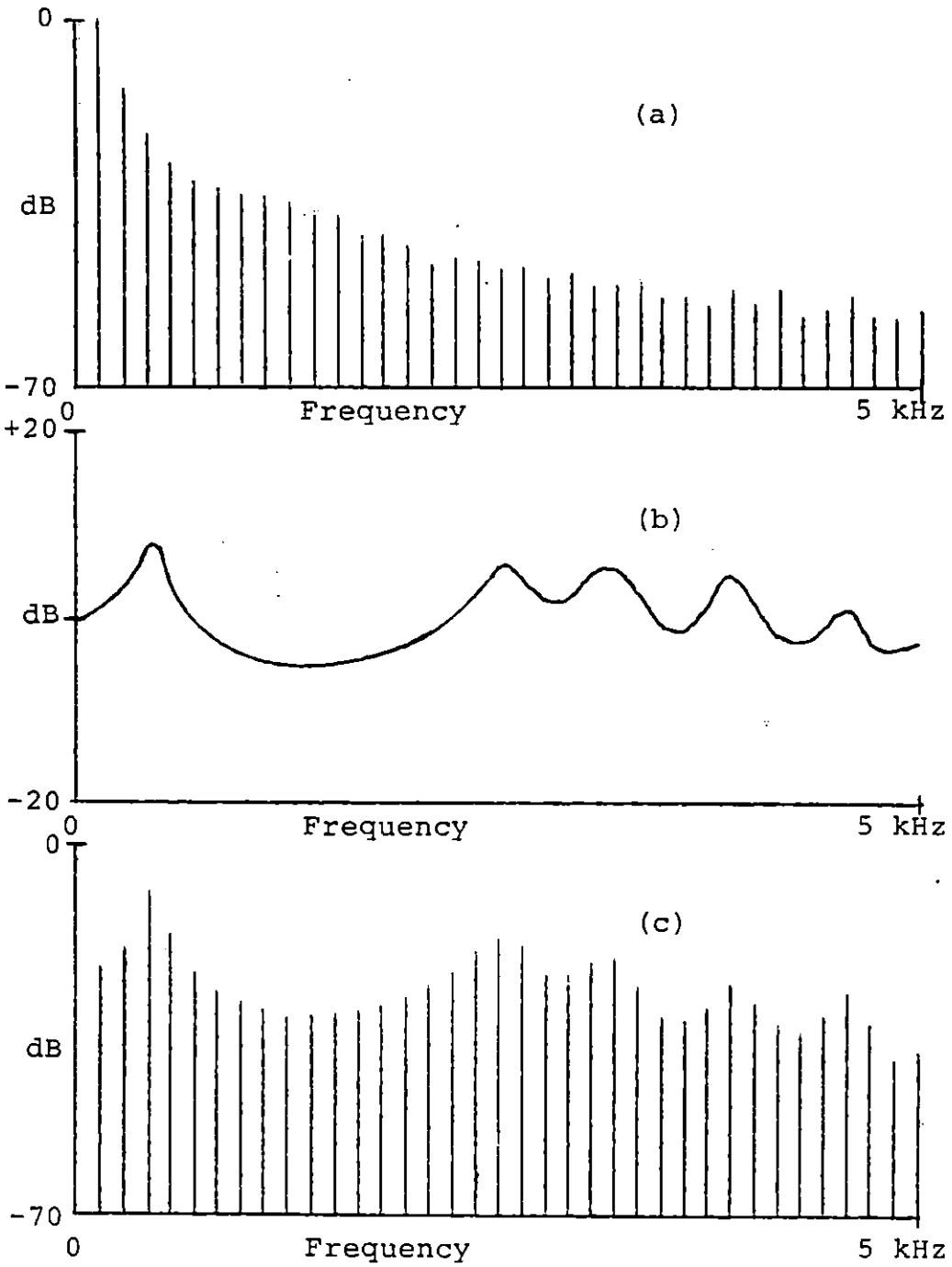

FIGURE 1. Typical speech characteristics. (a) Line spectrum of typical excitation. (b) Frequency response of the vocal tract. (c) Spectrum of resulting acoustical response (Holmes, 1972)

Implications

Clapper (1971) felt that the variations in amplitude of the original waveforms carry mainly speaker idiosyncrasies such as emotional state, age, gender, intonation, and other information peculiar to the speaker. It is certain that the properly preprocessed dichotomized speech waveforms are intelligible. That is, individual spoken words are recognizable with a high degree of accuracy. Licklider and Pollack (1948) found that up to 98% word recognition rates could be achieved for dichotomized speech.

The application of zero-crossing techniques to tactile hearing aids for the deaf, such as that discussed in O'Brien (1977) is of particular interest. His device has shown some success, but appears to be limited by the small bandwidth of the tactile sensory modality relative to the bandwidth of the acoustical signal. The device implemented in this thesis accomplishes much in the way of bandwidth reduction and therefore could perhaps be applied to tactile aids.

Goals

Figure 2 shows the locations of formant one and formant two frequencies for ten speech sounds. It is apparent that the locations of these formants may be very useful in distinguishing the phonemes.

A goal of this research was to examine the effectiveness of zero-crossing-rate analysis in determining formant locations. Once this was established, studies were done to determine whether these formant locations are sufficiently independent to allow classifications of speech sounds. The performance of different types of filters in effectively

isolating the first and second formants was evaluated. In addition, the system developed was to operate on-line to illustrate one of the primary advantages of zero-crossing-analysis, speed of operation.

Methods

In this research, the first and second formants were separated by suitably tuned bandpass filters. Once a satisfactory set of filters was developed, it was determined if this system could produce a sufficient separation of phonemes to allow one phoneme to be differentiated from another.

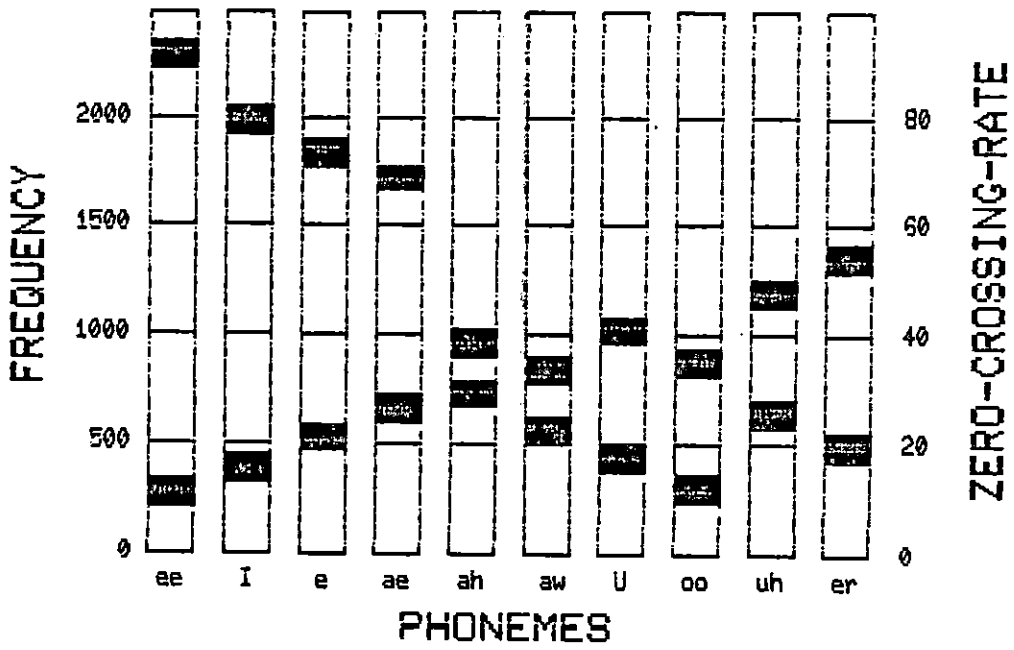

FIGURE 2. Frequencies of the first formant and the second formant for ten English vowels. Also indicated are the corresponding zero-crossing rates (Flanagan, 1965)

The number of zero-crossings in an interval of time was counted and used in histograms and scattergrams to study phoneme separation. One histogram of the zero-crossing-rates which occurred for the ten sounds was made for each filter. They were used to determine how closely the zero-crossing-rates approximate the formant frequencies. The scattergrams are plots of the number of zero-crossings from one filter versus the number of zero-crossings from the other filter. The scattergrams for a set of phonemes were superimposed to determine if they overlapped. How well the phonemes were separated was determined by the amount of overlap. Vowels were chosen because their frequency characteristics change more slowly than many consonants. The vowels in Table 1 were chosen to make a meaningful comparison to literature possible.

Results

The zero-crossing rates of the phonemes studied in this thesis were found to approximate the frequency of the first and second formants of the speech waveform. Separation of phonemes on the basis of their zero-crossing rates was shown to be possible. Correct classification of nine phonemes was approximately 78% based on the zero-crossing rates from two filters which selectively isolated the first and second formant ranges of the frequency spectrum.

Table 1. Test phonemes used in this thesis (adapted from Flanagan, 1965)

Phoneme	as used in
ah	father
oo	boot
ee	eve
I	bit
aw	call
uh	cup
U	foot
ae	cat
er	bird
e	met

LITERATURE REVIEW

Phoneme Recognition

The ideal speech recognition device should not be constrained by vocabulary size nor be speaker dependent. Because of the difficulties in achieving such a system, researchers have limited the vocabulary size and the number of speakers, and have thereby been able to generate acceptable results. Recognition has generally been accomplished by matching a temporal and/or spectral template of the words stored with the pattern of the word to be recognized. The limitations in both memory space and in speed of operation for large numbers of words are obvious. If storage and recognition of the entire English language, or even part of the English language leads to speed and memory space difficulties, a system which has a smaller number of units to recognize is desirable. Because individual sounds (phonemes) are the smallest units speech can be broken into (Lindgren, 1968), many researchers have attempted individual sound or phoneme recognition. Because there are only about 40 phonemes (the number varies depending on dialect) in the English language, the advantages in reduced memory space and in increased speed of operation are obvious.

Important Acoustical Properties

As mentioned before, it is generally felt that the speech sounds produced are determined by the configuration of the vocal tract and the source of excitation. Oppenheim (1970) felt that " . . . speech sounds may be considered to be produced by exciting a resonant cavity, the vocal

tract, with either a quasi-periodic or noiselike excitation . . . the resulting speech waveform is characterized by the frequencies of the vocal tract resonances . . . ." The excitation is due to vibration of the vocal cords or to turbulence caused by forcing air from the lungs through a constriction in the vocal tract. The differing responses of the of the vocal tract are due to motions of the tongue, lips, jaw, and velum. According to Holmes (1972), it is the motions of these articulating organs which determine the speech sounds produced, rather than the excitation sources. He indicates that the excitation affects mainly intonation. This is illustrated by Figure 1. Figure 1A shows the line spectrum of a typical excitation. The frequency response of the vocal tract is shown in Figure 1B and the spectrum of the resulting acoustical wave is shown in Figure 1C. It is apparent that the acoustical signal is the result of a shaping of the excitation by the vocal tract.

Apparently, the frequency response of the vocal tract is critical in the production of speech. The current assumption that the cochlea responds in some way to the component frequencies in the speech waveform (Kandel and Schwartz, 1981) indicates that information is carried by the frequencies present in the speech waveform.

If sounds are produced by changes in the resonant characteristics of the vocal tract, it is reasonable to assume that each sound has its own particular resonant qualities. Later discussions show that while some sounds may be characterized by individual resonant qualities, other sounds require additional information for characterization.

Researchers studied the spectra of speech sounds first with banks of bandpass filters, then later by linear prediction methods and the Fast Fourier Transform. The humps caused by the resonances of the vocal tract are obvious in the resulting frequency plots. These resonant frequencies, denoted formants, are shown in Figure 1C. Results obtained using formant frequencies to determine the sound spoken are encouraging. Bezdel and Chandler (1965) distinguished five sounds from each other with 88% accuracy using locations of the formant frequencies. Trunin-Donskoi and Tsemel (1968) found that by using the first and second formant frequencies to characterize the speech waveforms the following recognition rates resulted when these sounds were distinguished from each other: [o]-93.2%, [u]-70.4%, [a]-83%, [e]-74%, and [i,y]-87%. These and other representative studies have resulted in recognition rates of about 90% for similar groups of vowels. A system developed by Neiderjohn and Thomas (1973) recognizes all the continuant phonemes (ones which can be sustained) with a mean accuracy of 78% for a single speaker.

The above systems generally use a limited vocabulary. Most researchers use the continuant sounds because much of the information they contain is transmitted during the static portion of the waveform, making analysis simpler.

Systems which have similar recognition rates may be acceptable if one is interested in isolated word recognition of a limited vocabulary. However, assuming the ultimate goal of speech recognition is continuous speech recognition of a large vocabulary, it seems that more discrimination between sounds is desirable. Trunin-Donskoi and Tsemel (1968) discuss

this, stating their system provides enough discrimination between many sounds to be useful in characterizing them, but not enough discrimination to identify more than a few, especially for different speakers. Because of this, they feel their system may be more appropriate for a limited word recognition system than for phoneme recognition.

Limitations

Again, speech recognition based on formant information has provided encouraging results, but many difficulties are evident. Continuant sounds appear to be the easiest to identify. Well-designed systems which attempt to identify continuant sounds have percent recognition rates ranging from the mid-seventies to the mid-nineties, depending upon the phonemes chosen. In attempting to identify the continuant sounds, " . . . part of the difficulty appears to be a shift in formant frequencies of some words dependent upon the sounds adjacent to it. The shift is small but is sufficient to move a sound into another group" (Neiderjohn and Thomas, 1973).

Consonants have proven more difficult to categorize than vowels. "Experiments show that much information about the identity of a consonant is carried not by the spectral shape at the 'steady state' time of the consonants, but by its dynamic interactions with adjacent phonemes" (Neiderjohn and Thomas, 1973). So not only must the dynamic variations of the resonances of the consonants be analyzed, but so also must the way these variations vary depending upon which sounds precede and follow the particular consonant.

Speech sounds are not nicely packaged units as is pointed out by Lindgren (1968), who felt that phonemes are not invariant units, rather they flow from sound to sound. Phonemes merge and overlap, making feature extraction difficult.

Characterization of the transitions between, and the static properties of, phonemes is necessary for successful continuous speech recognition devices. In addition, semantic and syntactic information must be taken into account, for according to Nash-Weber (1975), "the acoustic signal we hear is so imprecise and ambiguous that even a knowledge of the vocabulary is insufficient to ensure correct understanding." This is also substantiated by Neiderjohn and Thomas (1973), who state, " . . . even the human perceptual system has difficulty recognizing certain speech sounds out of context." It appears that sequential integration of successive speech stimuli is essential to speech recognition, implying that the manner in which the frequency information varies as a function of time may be an important characteristic of the speech waveforms.

The general feeling as expressed by Lindgren (1968) is that continuous speech recognition will someday be possible, but not until we have a better idea how our mind perceives the speech signal. This does not mean that useful and effective systems cannot be developed prior to obtaining an exact model of the human auditory system. Systems have, in fact, been developed which work quite well for limited vocabularies and numbers of speakers, an example of which is described by Brumwell (1978).

Data Reduction by Infinite Clipping

It appears a crucial problem in continuous speech recognition is the reduction of the speech signal into a sufficiently informative but nonredundant set of features. The extreme range of amplitudes present requires system accommodation. "If speech is viewed in the time domain, an obvious characteristic is the great dynamic range of amplitudes which are present. Variations of up to 60 dB are not uncommon" (Ewing and Taylor, 1969). Licklider and Pollack (1948) have shown that amplitude information is nonessential. " . . . by clipping . . . the speech wave until it is reduced to a two-valued function of time, we can produce an intelligible temporal pattern." They found that the intelligibility of dichotomized speech is 86%. They went on to state "pre-emphasis of the high frequency components was found to increase intelligibility . . .up to 98% for trained listeners." This implies that higher frequency information is more important than lower frequency information for zero-crossing analysis. The intelligibility was determined for individual words. The intelligibility of connected speech would be even greater because of syntactic and semantic clues (Flanagan, 1965).

In the above case, data reduction is achieved by removing extraneous amplitude information leaving only the sequence of zero-crossings to convey information. It appears that most of the information lost is related to cues which express the "age, health, sex, emotional attitude, perhaps even place of birth and educational background of the speaker" (Clapper, 1971). It is doubtful whether the place of birth can be obtained from the speech waveforms, however one might assume that Clapper was referring to the

geographical location where one's speech idiosyncracies were formed. For these reasons, it is thought that infinitely clipped speech may be more appropriate than the original waveform in speaker independent systems (De Mori, 1971; Ito and Donaldson, 1971).

Zero-crossing Analysis of Speech

Peterson (1951) has shown that the number of zero-crossings in a band of frequencies from 200 to 1000 Hz is very close to the first formant frequency of a given speech sound and the number of zero-crossing in a band from 1000 to 4000 Hz is very close to the second formant of a given sound. If the sounds are assumed to be time invariant, it may be possible to extract formant information by suitably preprocessing the waveform to emphasize the frequency range of interest and then counting the axis crossings within a specified time period for which the signal is assumed to be unchanging. This has been tried, as have been methods involving a measurement of time durations between axis crossings (Neiderjohn and Thomas, 1973). Results have varied, but in general have been promising. Denes (1959) reports that 13 phonemes were recognized with a 70% recognition rate for one speaker. Forgie and Forgie (1959) report an impressive 90% recognition rate for 10 phonemes and 21 speakers, however, the speakers were "chosen", presumably based on their having acoustically similar voices. Neiderjohn and Thomas (1973) report recognizing 24 phonemes with a 78% recognition rate, however the phonemes recognized could generally be categorized by their time invariant or midsound properties, as opposed to their transient behavior. Also, only a single speaker was used.

Assumption of Time Invariance

It should be noted that in order that the rate of zero-crossings in an interval approximates the primary frequency present, that frequency should be unchanging. This assumption is realistic, for although the resonant properties of the system are time varying, they do change slowly enough that short duration windows of time may be taken to be time invariant (Oppenheim, 1970). Because vowels last at least 60mS, most researchers choose windows of 10-20mS. There is an essential trade-off here, for the larger the windows are made, the less the zero-crossing-rates reflect and depend upon transitions between phonemes. Such a situation is appropriate for vowel recognition since vowels may be characterized by their time invariant qualities, however, it may not be as suitable for some consonants, for it is these very transitions between the phonemes which contain the information about certain consonants. One must realize that in assuming the signal is quasi-static for small periods of time, and choosing a window size to reflect this, one loses the information about the sequence of intervals between crossings, and this information may be crucial, particularily in consonant identification, many of which depend more on changes in zero-crossing-rates than changes in steady-state or midsound properties (Neiderjohn and Thomas, 1973).

MATERIALS AND METHODS

Speech Analysis System

An on line system was developed to study the speech waveform. Two pairs of two parallel filters were used to study the frequency ranges of formant one and two. The filters used and the formants they emphasized are listed in Table 2.

The system consists of a microphone, an audio amplifier, two filter pairs (either F1A/F2A or F1B/F2B), two comparators to provide TTL compatible outputs, and two op amp buffer stages connected to a microcomputer system (Commodore Series 2001 Professional Computer, Commodore Business Machines, Inc., Santa Clara, Calif.). A block diagram of the system is shown in Figure 3.

The microphone has a flat frequency response (3 dB) to 9 KHz. The audio amplifier consists of LM324 Operational Amplifiers, with closed-loop gains small enough that the bandwidth remains greater than 10 KHz.

The audio amplifier output is applied to the inputs of the F1A/F2A or F1B/F2B filter pair. These filters were designed according to methods given in Johnson (1976) and Wait et al. (1975).

Filter F1A is a low-pass filter used to analyze formant one datum. Because the first formant is larger in amplitude than the second formant, it dominates the acoustic signal if the signal is viewed on an oscilloscope. The zero-crossing-rate of the original waveform should therefore be close to that of the first formant. Therefore, filter F1A was designed to have a flat frequency response in the range of the first and

second formants. A fourth-order low-pass Chebychev section with a break frequency of 2800 Hz was incorporated into the system to ensure that the sampling rate of 10 KHz was sufficient to register all the zero-crossings. This frequency is above the range of the second formant.

Filter F2A is a band-pass filter which emphasizes the second formant range. The band-pass filter consists of a low-pass stage similar to that mentioned above, cascaded with a second-order Butterworth high-pass filter which has a break frequency of 1000.

Filter F1B is a band-pass stage tuned to pass frequencies from 250 to 760 Hz. A fourth-order Chebychev high-pass filter reduces unnecessary energy present below 250 Hz. In Figure 2, the ranges of the first and second formants can be seen to overlap to a small degree. To effectively isolate the first and second formants, a sharp cutoff is required. The second formant amplitudes are smaller than the amplitudes of the first formant. This means that extreme attenuation is not required to eliminate the second formant. A fourth-order elliptic filter with a break frequency of 760 Hz was chosen because it met the above demands. Minimum attenuation of 22 dB was measured in the stopband of this filter (>960 Hz). This, combined with the hysteresis designed into the comparator, is sufficient to remove most effects of the second formant.

Filter F2B is a band-pass filter tuned to emphasize the formant two range. It consists of a fourth-order Chebychev low-pass filter with a cutoff of 2800 Hz. The high-pass section is a sixth-order Chebychev filter chosen for sharp cutoff. It has a cutoff of 830 Hz, just at the lower extremity of the range of the second formant for the sounds used.

The filter pair in the circuit is followed by a pair of comparators, one per stage. The comparators provide TTL compatable outputs, simplifying data acquisition by the computer. The author attempted to ensure that center clipping would not exceed 1 dB by adjusting the amount of hysteresis of the comparator. This was done because Licklider and Pollack (1948) have shown that center clipping of more than 1 or 2 dB drastically reduces the intelligibility of the speech waveform.

The IEEE bus of the Commodor Pet microcomputer was used to input data. A sampling program fetched the datum at a 10 KHz rate and also analyzed it for zero-crossings.

Experimental Methods

Table 1 shows the phonemes used in this thesis. The effect of each of the ten phonemes on the outputs of each of the four filter stages was studied. Histograms of the resulting zero-crossing-rates were plotted and compared to the predicted values. This illustrates how well the zero-crossing-rates of each filter approximated the formant it was designed to isolate. Sixty repetitions of each sound were plotted for each filter. Because the results appeared to be consistent, more points were thought unnecessary.

Next, one of the two filter pairs, F1A/F2A or F1B/F2B, was tested and the resulting zero-crossing-rates plotted in scattergram form. The variables were the zero-crossing-rates from one filter versus the zero-crossing-rates from the other filter of the pair. Each sound was repeated a total of 100 times for filter pair F1A/F2A, while 250

repetitions were made for F1B/F2B. More data points were recorded for filters F1B/F2B because this pair showed markedly better phoneme separation, and futher verification of this was desired.

Boundaries or decision surfaces were drawn to enclose as many of a given phoneme as possible. The boundaries did not overlap. Figure 8 shows the boundaries drawn from the points using the F1B/F2B filter pair. The data points are listed in Appendix B. Figure 9 shows the boundaries derived from the points obtained using the F1A/F2A filter pair. The data points obtained are listed in Appendix A.

After a choice of boundaries was made, they were tested. For the surfaces of Figure 8, each phoneme boundary was tested twice, with 100 repetitions of each sound the first time and 50 repetitions the second time. The phoneme boundaries shown in Figure 9 were tested once with 50 repetitions of each phoneme.

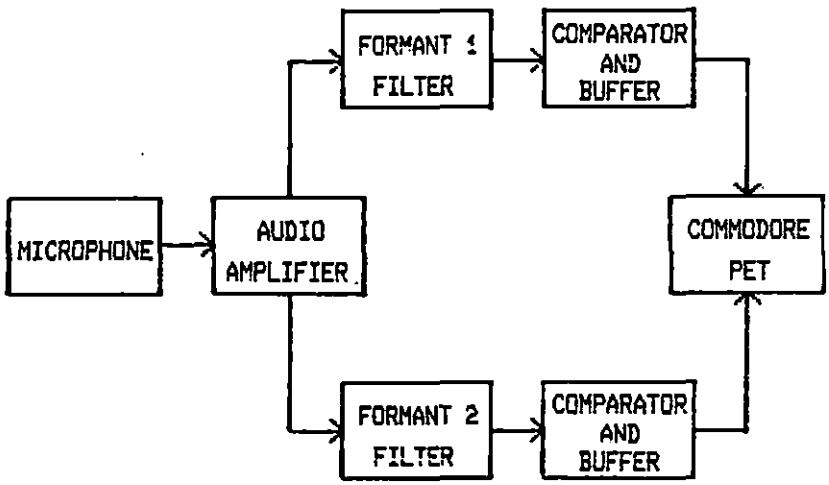

Figure 3. Block diagram of speech analysis system. Filters F1A and F1B are shown as the formant 1 filter. Filters F2A and F2B are shown as the formant 2 filter

Table 2. Summary of filters used

Label	Type of filter	Order	Break frequency (Hz)	Dominant formant
F1A	Chebychev low-pass	Fourth	2800	First
F2A	Chebychev low-pass	Fourth	2800	
	Butterworth high-pass	Second	1000	Second
F1B	Chebychev high-pass	Fourth	260	
	Elliptic low-pass	Fourth	760	First
F2B	Chebychev high-pass	Sixth	830	
	Chebychev low-pass	Fourth	2800	Second

RESULTS AND DISCUSSION

Introduction

As was mentioned in earlier sections, it is felt that the rate of zero-crossings of the speech waveform in a band from 200-1000 Hz approximates the rate which would occur if only the the first resonant response was present and similarily, the zero-crossing-rate in a band from 1000-4000 Hz approximates the rate which would occur if only the second resonant response was present (Peterson, 1951). Figure 2 shows the first and second formant frequencies for the phonemes which were chosen for testing. These are typical values and ranges as reported in Flanagan (1965) for a small group of male speakers. The zero-crossing-rates which would occur if only the resonant frequencies of the acoustical wave were present are also shown in Figure 2. The zero-crossing-rates from Figure 2 are superimposed upon Figures 4, 5, 6, and 7 for comparison purposes. Both the overlap which occurs, and the variance of 4-5 zero-crossings which could occur due to small changes in the formant frequencies are obvious. One should therefore not be surprised that the results displayed in Figures 4, 5, 6, and 7 are not centered at one zero-crossing-rate, but rather are spread, primarily, to 5 or 6 different rates. Another cause for a small amount of spread may be the sampling procedure itself. When one attempts to measure the frequency of a sinusoid, one might measure the time necessary for some number of zero-crossings and use this information to calculate the frequency. If one reverses the process and counts the zero-crossings in the waveform for a given amount of time, one does not

have the same degree of accuracy. The latter situation occurs in this thesis.

First Formant Analysis

The mean and range of the zero-crossing-rates from systems F1B and F1A are indicated in Figures 4 and 5, respectively. The results from system F1B follow the predicted trends relatively well, with the mean values being near the ranges obtained in Flanagan (1965), however some of the phonemes with lower resonant frequencies have slightly higher zero-crossing-rates than predicted. Because these sounds have resonances significantly below the filter cutoff, it is possible that energy in the band between the resonant frequency and the cutoff frequency could contribute to the zero-crossing-rates and thereby increase them. Figure 1C is the spectrum of a typical acoustical waveform and it shows there is significant energy on either side of the resonant peak. The results from filter F1A exhibit the same characteristic, in fact, the mean values for the phonemes in Figure 5 are even further from the predicted values, probably due to the additional higher frequencies retained in the waveform, giving support to the idea that the higher (supraresonant) frequencies in the passband of the filtered waveform cause the shift. Time functions of typical waveforms show humps which appear to be responsible for the higher zero-crossing rates. The humps appear to be of higher frequency than the larger amplitude resonant response.

Figure 5 shows a greater spread of zero-crossing-rates than Figure 4. This is most likely also caused by energy present above the first resonant

frequency. In this case, all of the formant two energy remains in the waveform, causing more spreading and a greater upward shift.

The results for system F1A are similar to that of F1B. Differences are generally what one would expect. Leaving the second formant in the waveform cause a larger number of zero-crossings, and because their effect is somewhat random, a greater variance of zero-crossing-rates is observed. Because discrimination between all ten phonemes is not evident in results from F1A or F1B, it may be that the effectiveness of the system is not increased significantly by the filtering process. For comparison purposes, from the histogram in Figures 4 it can be computed that [ee], [aw], and [I] can be separated into groupings which include 100% of [ee], 93% of [I], and 100% of [aw]. Similar groupings from Figure 5 would include 97% of [ee], 82% of [I], and 62% of [aw]. Similar comparisons for other groupings can be made and the results from filter F1B appears to be superior to that of F1A, although not necessarily to a great degree. As mentioned in the Materials and Methods section, filter F1B was chosen because it was thought to be sufficient to remove the second and subsequent formants from the speech waveform. Because of this, it is doubtful that further filtering could improve the results.

The results from system F1B are similar to those reported in Ito and Donaldson (1971). The recorded mean and range of the zero-crossing-rates are listed in Table 3. The mean values and ranges achieved for F1B (modified to a 10mS sampling period) are also shown in this table. No significant differences are present.

Second Formant Analysis

Two different filters (F2A and F2B) were used for isolation of the second formant. They are discussed in the Materials and Methods section. Licklider and Pollack (1948) have shown that the second formant is very important in the recognition of clipped speech. They found that emphasizing the second formant relative to the first resulted in more intelligible speech. Because of this, one might hope to see more phoneme separation for results from the F2A or F2B filter sections. Indeed, this is what occurred. The zero-crossing-rates for the output of filters F2B and F2A are shown in Figures 6 and 7, respectively. Separation of phonemes on the basis of the results in Figure 6, system F2B, results in significant separation for some phonemes. As an example of the increased discrimination possible with this filter, the zero-crossing-rates for seven phonemes were used to determine the percent which could be separated. An attempt was made to include the maximum number of phonemes possible in the seven groupings. The following percentages of sounds fell into non-overlapping groups: 100% of [ee], 92% of [I], 70% of [e], 85% of [er], 90% of [U], 78% of [oo], and 88% of [aw]. This is better than the separation seen in formant one datum. If the other three phonemes are included, separation drops significantly. A comparison of Figures 6 and 7 reveals that the separation which occurred in Figure 7, for system F2A, is not nearly so good as that in Figure 6, for system F2B. If one attempts to make similar groupings, from Figure 7, to above, one sees how significantly results are degraded.

Zero-crossing-rates for the output of filter F2A can be seen to be lower than that of filter F2B. This is thought to result from the increased effect of the first formant due to a filter with a larger transition band. The first formant is apparently large enough in amplitude to prevent some of the second formant oscillations from causing zero-crossings. This results in a lower and more variable zero-crossing-rate.

The mean zero-crossing-rates and the predicted values are labeled in Figures 6 and 7. Though the mean zero-crossing-rates are relatively close to the predicted values, there are some differences. A noticeable trend is present. The sounds with the higher second formants have lower rates than predicted. The sounds with lower second formants have higher rates than predicted, and the sounds with second formant frequencies between the two extremes match the predicted values rather well. Refer again to Figure 1C, the spectrum of a typical acoustic waveform. If the location of the second formant is visualized as moving to higher or to lower frequencies, the subresonant frequency energy can be seen to increase or decrease while the supraresonant frequency energy does the converse. There is a intermediate region where these effects cancel. As the location of the resonant peak is moved higher, more low frequency energy is added to the signal, decreasing the zero-crossings and vice versa.

As seen in Figures 6 and 7, the sound [ae] exhibits a downward shift of much greater magnitude than the other sounds. This extra shift may be caused because the first formant peak has not reached a minimum before the passband of the high-pass filter begins. This would leave additional

Table 3. Comparison of zero-crossing-rates of speech waveforms preprocessed with filter F1B versus those achieved by Ito and Donaldson (1971)

	F2B zero-crossing-rates (for 10mS)		Literature zero-crossing-rates (for 10mS)	
Sound	Mean	Range	Mean	Range
ee	7	6-8	6	5-7
I	9.5	8-12	10	9-12
e	11.5	11-13	11	11-13
ae	14	13-15	14	13-17
oo	7.7	7-9	7	6-8
U	11	9-13	10	10-11
aw	12.7	12-15	15	13-17
uh	13	11-14	14	12-16

Table 4. Comparison of zero-crossing-rates of speech waveforms preprocessed with filter F2B versus that achieved by Ito and Donaldson (1971)

	F2B zero-crossing-rates (for 10mS)		Literature zero-crossing-rates (for 10mS)	
Sound	Mean	Range	Mean	Range
ee	42	38-44	50	45-56
I	35	32-37	41	33-49
e	31	26-35	37	32-42
ae	22	17-30	36	32-50
oo	20	18-24	21	15-30
U	24	22-27	21	20-21
aw	22	21-25	22	19-22
uh	25	23-29	28	25-32

energy in the low frequency region. Ito and Donaldson's (1971) results do not show this shift, most probably because their filter cutoff is placed at 1000 Hz, and may be sharp enough to remove more of the first formant energy.

A comparison of the results from system F2B to those reported in Ito and Donaldson (1971) is shown in Table 4. The results achieved in this thesis are of somewhat lower frequency, again, probably due to the choice of filter passbands. The author's choice of a lower frequency cutoff left more low frequency energy in the signal. The ranges of zero-crossing-rates are also significantly lower for system F2B than Ito and Donaldson's. They may not have used a filter with a sharp enough cutoff, thus causing results which were less consistent. It may also be that by choosing 1000 Hz as a cutoff, they removed to much of the information from sounds with second formants below 1000 Hz. It is also possible that differences in speakers were significant.

Two Dimensional Phoneme Separation

Because the system discussed in this thesis did show some separation of phonemes based on the zero-crossing-rates of the preprocessed waveform (as discussed in the two previous sections), one would expect that using both formant one and formant two data simultaneously might allow for more effective separation of the phonemes. For example, based on the formant two information in Figure 6 [oo] and [ee] are completely separated while based on formant one information from Figure 4, they overlap. The phonemes [oo] and [a] overlap in Figure 6, yet are completely separated in Figure 4. In these cases, utilizing both formant one and formant two data results in

nearly 100% separation between sounds which could not be as effectively separated on the basis of either the first (Figure 4) or the second (Figure 6) formant alone.

Figures 8 and 9 show boundaries made between adjacent phonemes. The phoneme [ae] is not shown, but will be discussed later. Simultaneous determinations of the zero-crossing-rates from each filter system were made and the resulting rates plotted in a scattergram of formant one versus formant two. As discussed earlier, nonoverlapping boundaries between the phonemes shown on the scattergrams were drawn to include as many of one type of phoneme as possible. There is no claim that the resulting decision surfaces are the optimum choices.

Zero-crossing analysis of the outputs of filters F1B/F2B resulted in the data points listed in Appendix B. The boundaries established from these points are shown in Figure 8. After choices of phoneme boundaries were made, they were tested. Table 5 shows the percentage the first 100 repetitions of each phoneme fell within its decision surface, and the number of times a particular phoneme fell within its decision surface for the subsequent trials. Data points obtained from filter pair F1A/F2A are listed in Appendix A and were used to establish the boundaries shown in Figure 9. Table 5 also shows the percentages of the data points which fell within the decision surfaces when they were developed and also the results of a test of the decision surfaces.

It may be that the percentage each phoneme was correctly categorized could be increased by a more sophisticated pattern recognition process, however even with the simple borders used, Table 5 shows that about 78% of

the phonemes fell within their respective boundaries for the results obtained utilizing the F1B/F2B filter pair. Only in two instances did a significant difference occur in the number of phonemes which fell within the selected borders occur for separate trials. These two were probably due to small changes in the way the sound was repeated, slightly changing the number of zero-crossings.

The outputs of the F1A/F2A filter pair show less separation than that of the F1B/F2B filter pair. This was predictable because of the greater spread of zero-crossing-rates which are exhibited in Figures 6 and 7. It was very difficult to draw boundaries between phonemes because of the large amount of overlap. Boundaries were finally drawn which enclosed 66% of the phonemes. Results for each phoneme are again listed in Table 5. Testing of the chosen boundaries resulted in only a 50% recognition rate. More significantly, six of the nine phonemes were recognized 20% less than the percentage which were enclosed when making the boundaries. Although this may point to a need for a better method of determining boundaries, this was deemed unnecessary because the data were obviously less consistent and structured than that of the F1B/F2B filter system.

Peterson and Barney (1952) and Foulkes (1961) both used spectral analysis techniques to determine the locations of the fundamental and the first three formants. They both recognized ten vowels with about 90% accuracy. Forgie and Forgie (1959) also used spectral techniques to locate the first two formants. They recognized ten vowels with an accuracy of 88%. Spectral analysis is usually done off-line with a spectrograph. The system implemented in this thesis, even using the simple boundaries,

achieved a separation of about 78% for nine vowels and allowed on-line analysis.

Sacai and Inoue (1960), Bezdel and Chandler (1965), and Trunin-Donskoi and Tsemel (1968) all analyzed sets of 5 vowels by their zero-crossing-rates. Sacai and Inoue correctly recognized 88%, Bezdel and Chandler correctly recognized 88%, and Trunin-Donskoi and Tsemel correctly recognized 81%. If the six phonemes [oo], [U], [er], [I], [ee], and [a] were the test phonemes, the authors system would have correctly classified about 96% of these phonemes. However if the phonemes [aw], [uh], [e], [I], and [U] were chosen correct classification drops to about 73%. One can see that the choice of phonemes makes a great deal of difference in system performance, making comparisons difficult.

Neiderjohn and Thomas (1973) have been very successful in phoneme recognition of clipped speech. They have developed a system which can identify 24 phonemes. The percentages they achieved for the phonemes used in this thesis are shown in Table 5. As can be seen, the overall recognition rates are very similiar to those obtained by the author utilizing filters F1B/F2B. Differences in individual phonemes may be due to choice of decision surfaces. Their results are for a larger number of phonemes than that used in this thesis. This required them to make additional measurements. They used the outputs of five filters and time as their variables and were thus able to get 78% recognition of 24 phonemes. The circuit implemented in this thesis used only two filters and correctly classified about 78% of nine phonemes. Separation of [uh] and [aw] illustrates the benefits of an additional dimension. The phonemes [uh] and

[aw] overlap a great deal, but because [uh] is a short vowel and [aw] is a long vowel, if time is added as an additional measurement, these two phonemes will exhibit a much greater degree of separation than otherwise seen.

The vowel [ae] was used in the initial testing of the performance of the system. Because the results for this vowel were inconsistent with expected results, as explained earlier in the discussion, and because the author found himself unable to establish appropriate boundaries, it was deleted from the tests of system performance.

Table 5. Results of system performance and comparison to the performance achieved by Neiderjohn and Thomas (1973) for a comparable group of phonemes

	F1B/F2B filter system			F1A/F2A filter system		Literature values
Phoneme	% within original borders	results trial 1 %	results trial 2 %	% within original borders	results trial 1 %	Neiderjohn and Thomas %
ah	82	79	96	64	62	80
oo	96	95	100	100	100	100
U	83	76	76	30	10	74
er	86	77	88	78	58	-
ee	95	90	91	100	96	82
I	91	63	82	84	62	62
e	54	66	54	48	22	89
uh	65	49	41	42	16	71
aw	82	76	76	52	28	-
Average	82	75	78	66	52	79

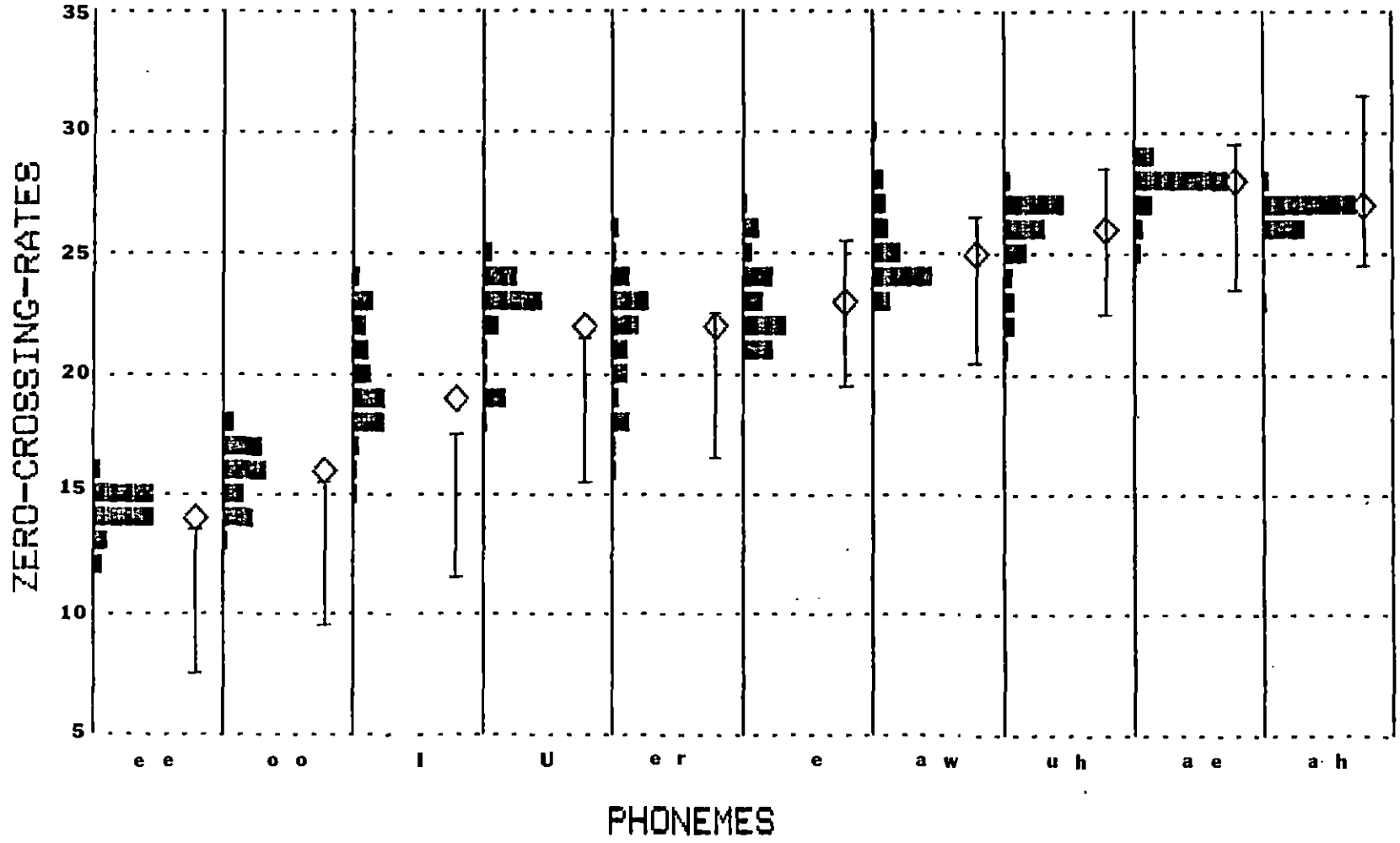

Figure 4. Zero-crossing-rates which resulted from tests of system F1B for each of the phonemes used. The diamonds represent the mean values of the zero-crossings. Corresponding literature values, from Figure 2, are shown as horizontal lines to the right of the associated data

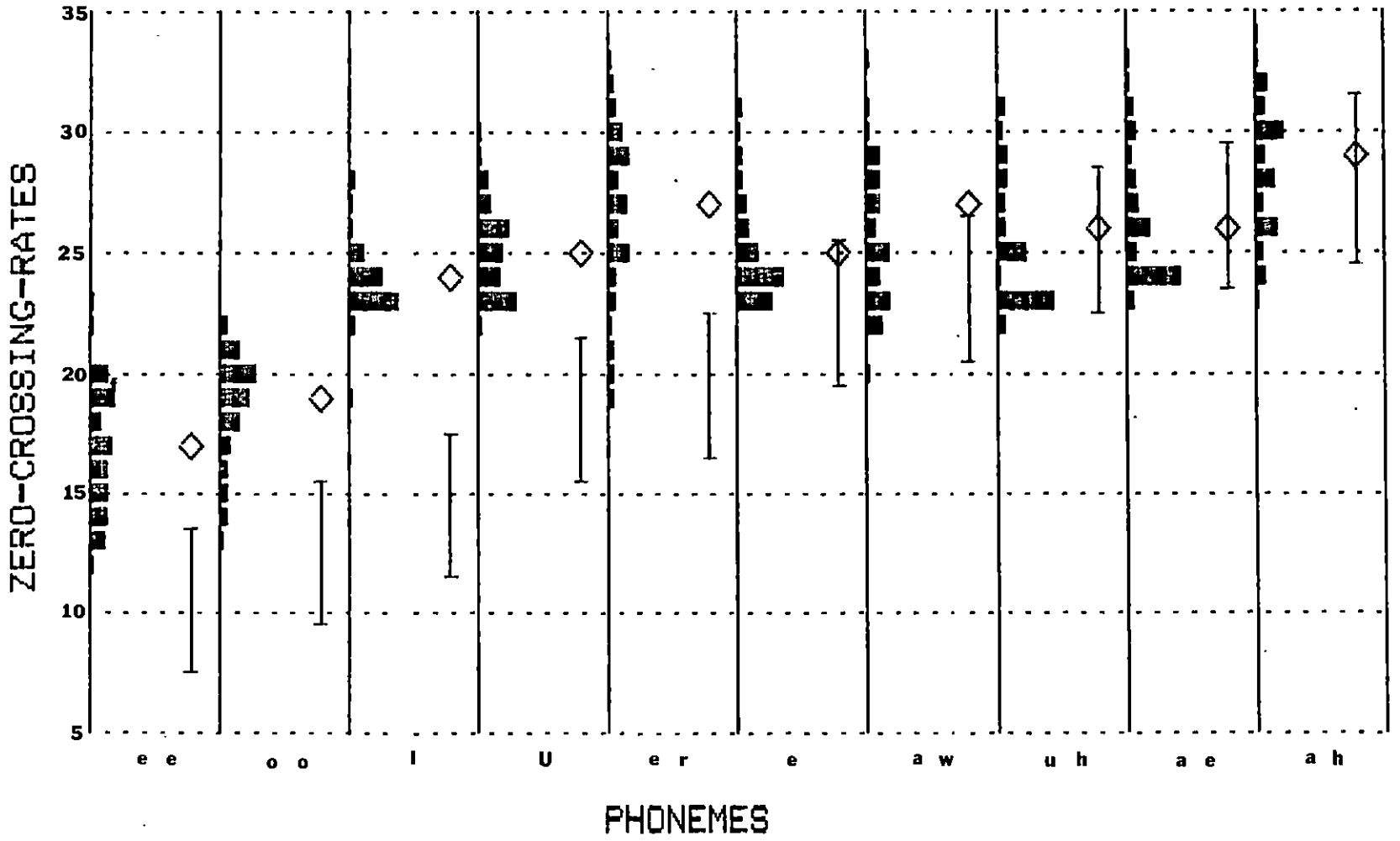

Figure 5. Zero-crossing-rates which resulted from tests of system F1A for each of the phonemes used. The diamonds represent the mean values of the zero-crossings. Corresponding literature values, from Figure 2, are shown as horizontal lines to the right of the associated data

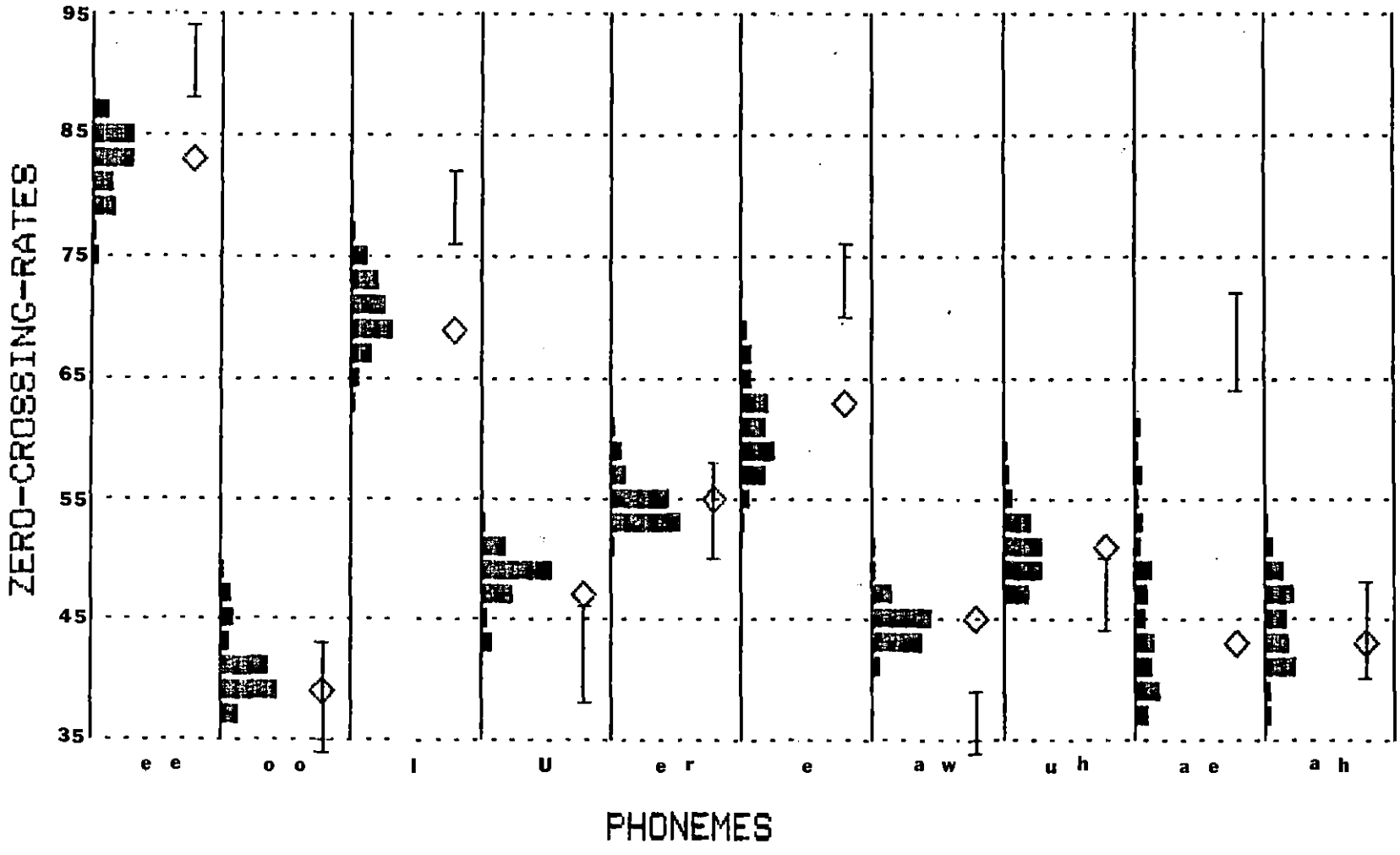

Figure 6. Zero-crossing-rates which resulted from tests of system F2B for each of the phonemes used. The diamonds represent the mean values of the zero-crossings. Corresponding literature values, from Figure 2, are shown as horizontal lines to the right of the associated data

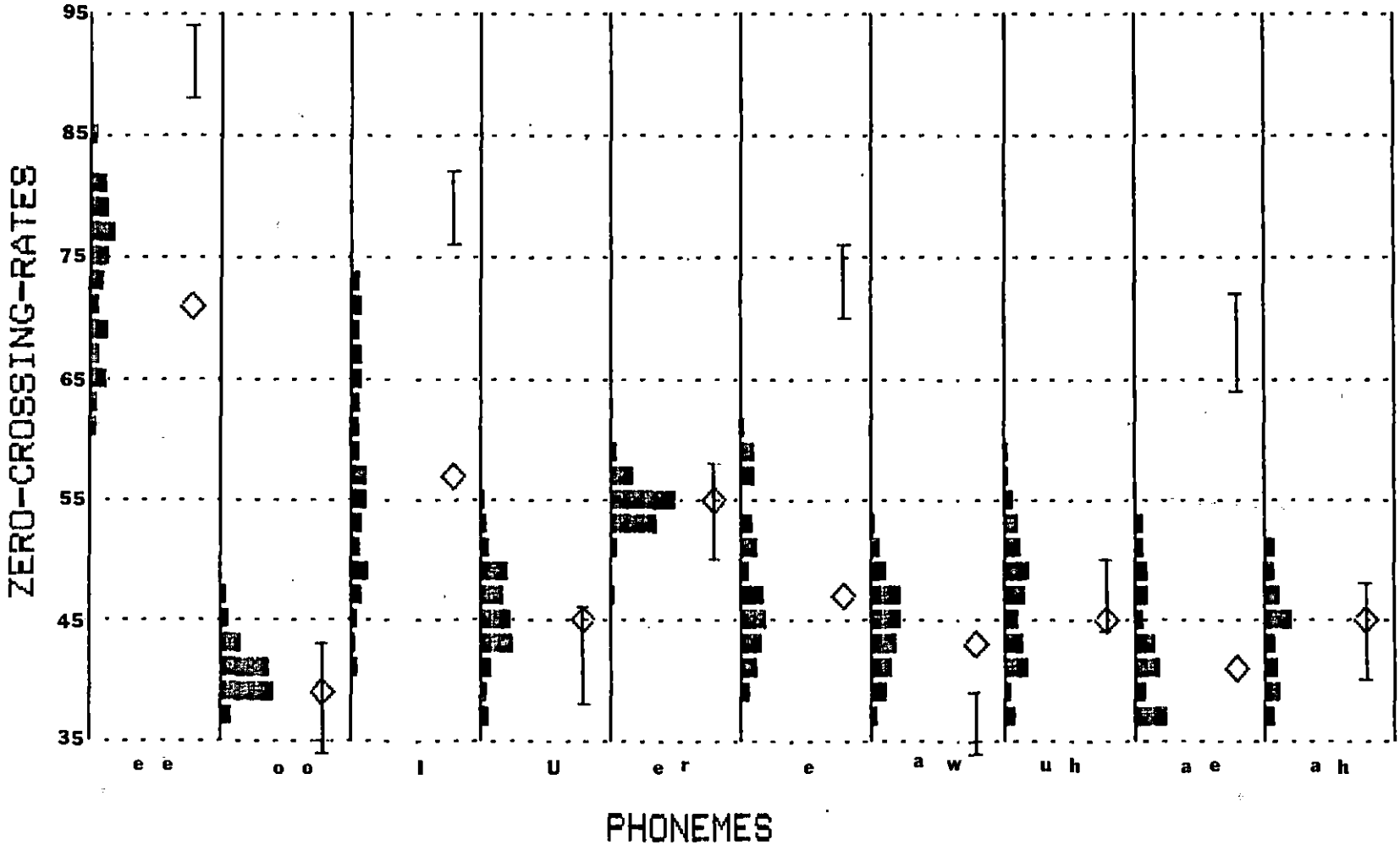

Figure 7. Zero-crossing-rates which resulted from tests of system F2A for each of the phonemes used. The diamonds represent the mean values of the zero-crossings. Corresponding literature values, from Figure 2, are shown as horizontal lines to the right of the associated data

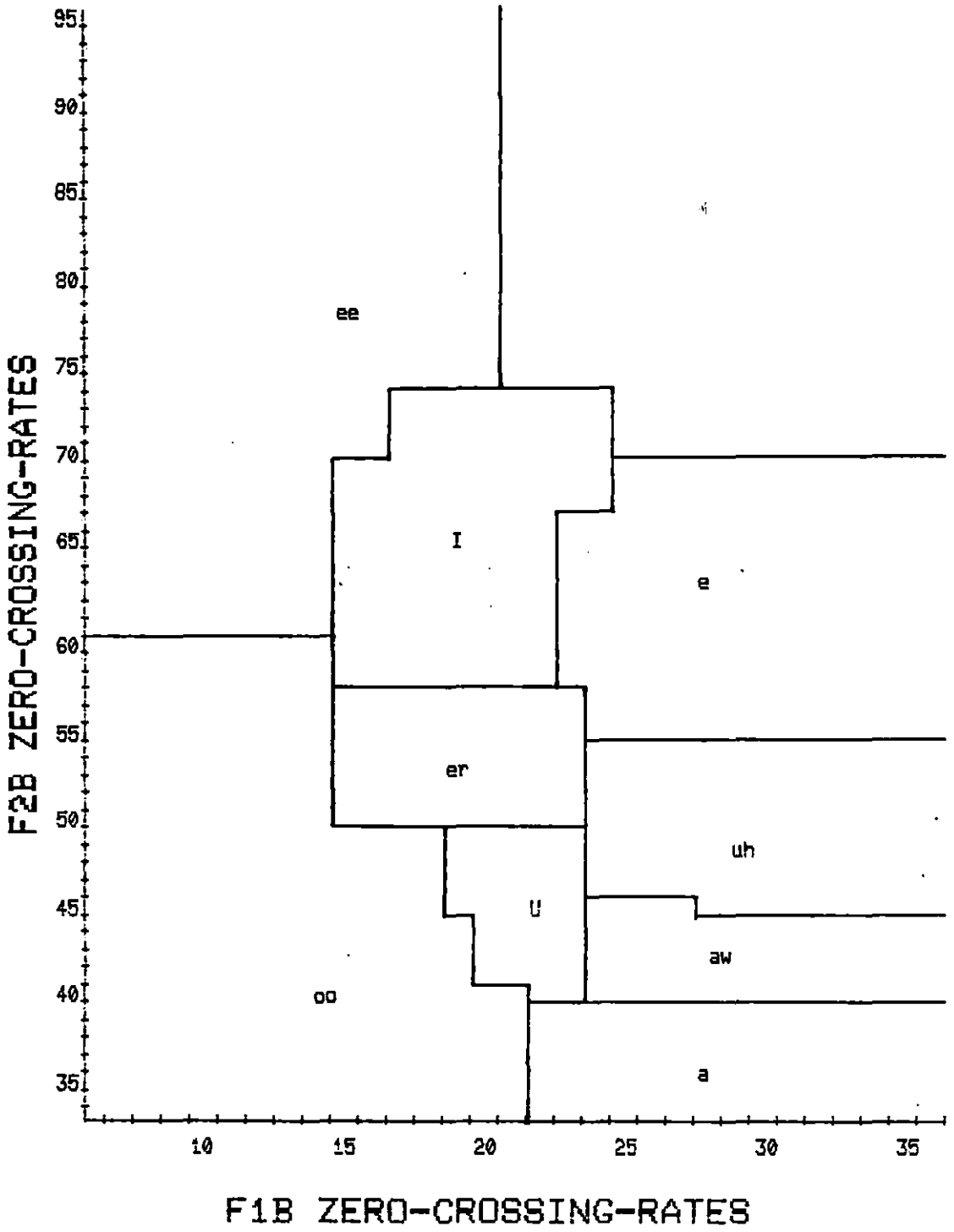

Figure 8. Decision surfaces which resulted in 75-82% correct phoneme classification for the filter system F1B/F2B

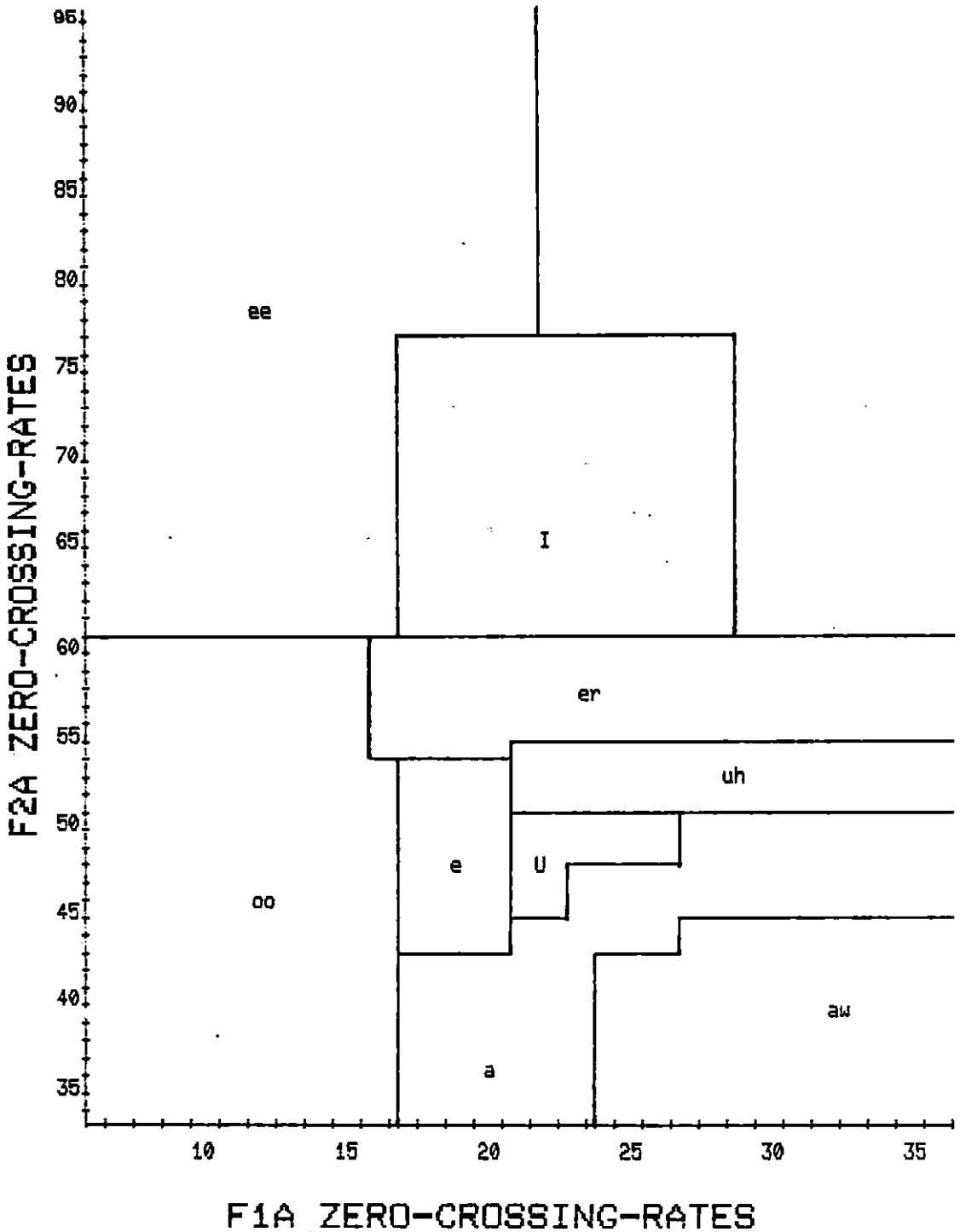

Figure 9. Decision surfaces which resulted in 50-66% correct phoneme classification for the filter system F1A/F2A

CONCLUSION AND RECOMMENDATIONS

A system was developed which could classify nine phonemes based on the zero-crossing-rates of the output of a pair of filters which emphasized the first and second formants of the acoustic signal. The overall classification rate was 78%.

It was shown that the original unfiltered (in the range of formant one and two) speech waveform has a zero-crossing-rate of approximately the first formant frequency. It was also shown that a sharper low-pass filter, with a break frequency located at 760 Hz, gives more consistent zero-crossing-rates.

A second-order high-pass filter, located at 1000 Hz, used to emphasize the second formant frequency range resulted in a zero crossing rate near that of the second formant, however a sixth-order high-pass filter, with a break frequency of 830 Hz, was shown to give more consistent and accurate results.

Because the computations involved in calculating the zero-crossing-rates are relatively simple, it may be possible to implement the phoneme recognition process in real time. The limiting factor would be the complexity of the recognition program. This system did allow for an on-line study of the speech waveform.

Licklider and Pollack (1948) demonstrated that properly preprocessed infinitely clipped speech is 98% intelligible. Because most of the information necessary for recognition of speech is present in the clipped speech signal, it should be possible to generate results as good or better than those achieved in this thesis by zero crossing analysis of the

original unfiltered waveform. When determining the number of crossings within a small time period, as done in this research, the sequence of these crossings is lost. The sequence of zero-crossings is related to the changes in the resonant response of the vocal tract and it may be important. A technique known as Walsh waveform analysis is particularly suited to digital signals and may enable researchers to study the clipped speech waveform without loss of the sequence of the zero-crossings.

Speech center clipped more than 1 or 2 dB has been shown to be relatively unintelligible (Licklider and Pollack, 1948). This is probably due to the loss of the smaller magnitude second and higher formant information the center clipping causes. Because this system separates formant one and two information before clipping, it should be relatively insensitive to center clipping (provided the center clipping does not remove the primary frequency passed by the filter system). It would be possible to verify this by comparing the results which could be obtained from a similar system with variable amounts of center clipping.

REFERENCES

Bezdel, W. and Chandler, H. J. 1965. Results of an analysis and recognition of vowels by computer using zero-crossing data. Proc. IEEE 112(11):2060-2066.

Brumwell, D. A. 1978. An isolated-word speech recognition system. M.S. Thesis. Iowa State University, Ames, Ia. 57 pp.

Clapper, G. L. 1971. Automatic word recognition. IEEE Spectrum 8(8):57-63.

De Mori, R. 1971. Speech analysis and recognition by computer using zero-crossing information. Acustica 25:269-279.

Denes, P. 1959. The design and operation of the mechanical speech recognizer. J. Br. IRE 19(4):219-229.

Denes, P. B. and Pinson, E. N. 1963. The speech chain. Waverly Press, Inc., Baltimore, Maryland. 158 pp.

Ewing, G. D. and Taylor, I. F. 1969. Computer recognition of speech using zero-crossing information. IEEE Trans. Audio Electroacoust. AU-17:37-40.

Flanagan, J. L. 1965. Speech analysis and perception. Academic Press [AInc., New York. 317 pp.

Forgie, J. W. and Forgie, C. D. 1959. Results obtained from a vowel recognition computer program. J. Acoust. Soc. Am. 31(11):1480-1489.

Foulkes, J. D. 1961. Computer identification of vowel types. J. Acoust. Soc. Amer. 33(1):7-11.

Holmes, J. 1972. Speech systhesis. Mills and Boon Limited, London, England. 64 pp.

Ito, M. R. and Donaldson, R. W. 1971. Zero-crossing measurements for analysis and recognition of speech sounds. IEEE Trans. Audio Electroacoust. AU-19:235-242.

Johnson, D. E. 1976. Introduction to Filter Theory. Prentice-Hall, Inc., Englewood Cliffs, New Jersey. 307 pp.

Kandel, E. R. and Schwartz, J. H. 1981. Principles of Neural Science. Elsevier North Holland, Inc., New York. 731 pp.

Licklider, J. R. and Pollack, I. 1948. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. J. Acoust. Soc. Am. 20:42-51.

Lindgren, N. 1968. Directions for speech research. IEEE Spectrum 5(3):83-88.

Nash-Weber, B. 1975. Semantic support for a speech understand system. IEEE Trans. on Acoustics, Speech, and Signal Processing 23(1):124-128.

Neiderjohn, R. J. and Thomas, I. B. 1973. Computer recognition of the continuant phonemes in connected English speech. IEEE Trans. Audio Electroacoust. AU-21:526-535.

O'Brien, E. M. 1977. A tactile hearing aid based on the properties of amplitude- and time-quantized speech. M.S. Thesis. Iowa State University, Ames, Ia. 138 pp.

Oppenheim, A. V. 1970. Speech spectrograms using the fast Fourier transform. IEEE Spectrum 17(8):57-62.

Peterson, E. 1951. Frequency detection and speech formants. J. Acoust. Soc. Am. 23(6):668-674.

Peterson, E. and Barney, H. L. 1952. Control methods used in the study of the vowels. J. Acoust. Soc. Am. 24(2):175-184.

Sacai, T. and Inoue, S. 1960. New instruments and methods for speech analysis. J. Acoust. Soc. Amer. 32(4):441-450.

Thomas, I. B. 1968. The influence of first and second formants of the intelligibility of clipped speech. J. Audio Eng. Soc. 16(2):182-185.

Trunin-Donskoi, V. N. and Tsemel, G. I. 1968. Recognition of vowel sounds from a clipped speech signal. Probl. Inf. Trans. (USA). 4(2):62-71.

Wait, J. V., Huelsman, L. P., and Korn, G. A. 1975. Introduction to Operational Amplifier Theory and Applications. McGraw-Hill, Inc., New York. 396 pp.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. William H. Brockman for all the help and advice he has given me. I would like to thank Dr. Charles Townsend for serving on my committee and also for making my introduction into teaching a rewarding experience. I am grateful to Dr. David Carlson for agreeing to serve on my committee. I am especially thankful to both the Biomedical and Electrical Engineering Deptartments for the financial support which allowed me to continue my education.

APPENDIX A

Data points used to develop the boundaries in Figure 9. F1 refers to the F1A system results while F2 refers to the F2A system results.

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
ee	10	68	I	13	70	e	15	43
	10	79		15	62		15	48
	11	74		17	58		17	43
	11	81		17	59		17	44
	12	60		17	63		17	46
	12	67		17	65		17	47
	12	69		17	69		17	48
	12	75		17	71		17	48
	12	77		17	71		17	52
	13	62		17	72		17	55
	13	66		17	72		18	46
	13	71		17	72		18	49
	13	72		17	72		18	50
	13	74		17	73		18	51
	13	75		18	57		18	51
	13	78		18	66		19	44
	13	78		18	67		19	45
	13	78		18	67		19	46
	13	78		18	67		19	46
	13	79		16	69		19	47
	13	81		18	70		19	49
	14	77		18	75		19	49
	14	77		19	64		19	49
	14	77		19	66		19	51
	14	78		19	66		19	53
	14	78		19	67		19	55
	14	79		19	70		19	56
	14	80		19	71		20	50
	14	81		19	71		20	51
	14	85		19	71		20	52
	14	86		19	72		21	42
	15	64		19	72		21	43
	15	69		19	73		21	52
	15	76		20	60		21	53
	15	76		20	61		22	49
	15	77		20	67		22	50
	15	79		20	68		22	54
	15	80		20	68		22	55
	15	80		20	69		23	47
	16	68		20	73		23	47
	16	72		21	67		23	48
	16	79		21	68		23	48
	16	82		21	70		23	51

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
ee	17	82	I	21	70	e	24	48
	18	78		21	73		24	49
	18	82		21	50		25	47
	18	83		21	50		26	45
	19	83		24	65			
er	16	54	U	17	47	u	11	40
	16	55		17	57		11	42
	16	55		18	47		11	45
	17	52		18	56		11	48
	17	54		19	50		11	50
	17	56		19	57		12	40
	17	59		20	51		22	42
	18	54		20	52		13	39
	18	54		20	53		13	40
	18	54		20	54		13	40
	19	53		21	45		13	41
	19	53		21	45		13	41
	19	53		21	47		13	43
	19	53		21	56		13	43
	19	53		21	58		13	44
	19	54		21	58		13	46
	19	55		22	40		13	48
	19	59		22	43		14	38
	20	52		22	46		14	39
	20	54		22	47		14	39
	20	54		22	48		14	40
	21	51		22	53		14	41
	21	54		22	55		14	41
	21	54		22	55		14	41
	21	55		23	53		14	42
	21	56		24	46		14	42
	21	56		24	47		14	43
	22	54		24	48		14	43
	23	53		24	48		14	43
	23	56		24	49		14	43
	24	55		24	49		14	44
	24	55		24	51		14	45
	25	47		24	52		14	46
	25	49		24	52		14	47
	25	54		24	55		15	39
	25	54		24	55		15	39
	25	54		25	45		15	39
	25	54		25	47		15	41
	25	55		25	47		15	45
	25	57		25	49		15	46
	26	52		25	54		15	47
	26	54		25	58		15	48
	26	56		26	47		15	50

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
er	27	55	U	26	53	u	15	51
	28	53		26	55		16	45
	28	55		27	57		16	48
	31	51		28	48		13	55
	32	56		28	52		16	54
	26	63		28	57		16	58
uh	18	43	aw	18	45	a	22	41
	19	53		21	37		22	42
	20	39		22	37		23	38
	20	54		22	41		23	39
	21	46		22	47		23	39
	21	51		23	37		23	40
	22	51		23	40		23	40
	22	53		23	41		23	40
	22	53		23	44		23	41
	23	41		23	44		23	43
	23	44		23	45		23	44
	23	46		23	46		23	46
	23	47		23	46		23	47
	23	48		23	46		23	48
	23	51		23	46		24	42
	23	51		23	48		24	42
	24	43		23	50		24	43
	24	49		24	40		24	43
	24	50		24	41		24	43
	24	51		24	44		24	44
	24	51		24	47		24	44
	24	51		25	40		24	45
	24	52		25	41		24	46
	24	52		25	44		24	46
	24	52		25	44		25	40
	24	53		25	45		25	44
	24	53		26	38		25	46
	24	54		26	41		25	46
	24	54		26	41		25	48
	24	54		26	49		25	49
	24	55		27	34		26	37
	24	56		27	35		26	40
	24	57		27	39		26	40
	24	58		27	41		26	41
	24	58		27	42		26	46
	25	43		27	43		26	50
	25	52		27	46		26	58
	25	55		28	34		27	37
	25	56		28	38		27	48
	26	43		28	41		27	48
	26	45		28	43		28	39
	26	48		29	35		28	39

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
uh	26	49	aw	29	39	a	28	41
	26	50		29	40		28	43
	26	52		29	40		28	48
	26	53		29	50		28	48
	26	54		29	51		29	43
	26	56		29	51		29	48
	27	52		29	51		30	43
							30	50

APPENDIX B

Data points used to determine the boundaries of Figure 8. F1 refers results from system F1B while F2 refers to results from system F2B.

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
ee	10	82	I	15	67	e	20	54
	11	72		16	64		20	67
	11	79		16	64		20	69
	11	84		16	68		21	49
	12	67		17	64		21	64
	12	70		17	65		21	65
	12	76		17	69		22	43
	12	78		17	69		22	45
	12	80		17	70		22	45
	12	81		17	71		22	50
	13	58		18	59		22	50
	13	63		18	59		22	52
	13	64		18	62		22	54
	13	64		18	63		22	55
	13	65		18	63		22	56
	13	65		18	64		22	56
	13	67		18	67		22	57
	13	69		18	68		22	58
	13	69		18	68		22	58
	13	69		18	68		22	59
	13	70		18	68		22	60
	13	70		18	70		22	62
	13	70		18	72		22	62
	13	71		18	72		22	62
	13	72		19	59		22	64
	13	72		19	62		22	66
	13	73		19	62		23	43
	13	73		19	62		23	46
	13	73		19	63		23	46
	13	73		19	63		23	48
	13	75		19	64		23	50
	13	76		19	64		23	50
	13	77		19	64		23	53
	13	78		19	64		23	56
	13	79		19	64		23	57
	13	79		19	65		23	58
	13	80		19	65		23	58
	13	80		19	66		23	59
	13	80		19	66		23	59
	13	81		19	67		23	62
	13	81		19	67		23	63
	13	81		19	67		23	64
	13	81		19	68		23	64

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
ee	13	82	I	19	68	e	23	64
	13	84		19	69		23	65
	13	86		19	70		23	65
	13	87		19	73		23	65
	13	89		20	55		23	66
	14	63		20	58		23	66
	14	64		20	62		24	62
	14	66		20	62		24	62
	14	67		20	65		24	65
	14	70		20	66		24	66
	14	72		20	66		25	43
	14	72		20	66		25	47
	14	76		20	66		25	48
	14	76		20	67		25	49
	14	76		20	68		25	50
	14	77		20	70		25	53
	14	78		20	72		25	53
	14	78		21	58		25	53
	14	79		21	59		25	56
	14	80		21	62		25	57
	14	81		21	64		25	60
	14	82		21	64		25	60
	14	83		21	64		25	60
	14	85		21	65		25	61
	14	86		21	65		25	61
	14	86		21	65		25	61
	14	88		21	66		25	62
	15	63		21	66		25	64
	15	65		21	66		25	65
	15	74		21	66		26	46
	15	77		21	67		26	49
	15	80		21	68		26	55
	15	85		21	69		26	57
	16	64		21	69		26	57
	16	70		21	69		26	58
	16	74		21	70		26	59
	16	77		21	70		26	59
	16	77		21	70		26	59
	16	82		21	71		26	59
	16	83		22	59		26	59
	17	70		22	66		26	60
	17	75		23	63		26	60
	17	78		23	64		26	61
	17	79		23	64		26	63
	17	80		23	67		26	66
	17	84		23	68		26	68
	18	76		23	68		27	49
	18	76		23	69		27	58
	18	78		24	63		27	62

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
ee	18	80	I	24	64	e	27	62
	18	81		24	66		27	64
	18	84		25	58		27	65
	18	85					27	66
	19	78						
	19	80						
	19	83						
er	15	55	U	18	42	u	10	40
	16	56		19	42		13	40
	17	56		19	46		13	41
	18	52		19	46		14	40
	18	53		19	48		14	40
	19	52		19	48		14	44
	19	52		19	50		14	47
	19	53		19	52		15	40
	19	53		20	41		15	40
	19	53		20	41		15	40
	19	54		20	41		15	40
	19	54		20	41		15	40
	19	55		20	43		15	40
	19	56		20	44		15	42
	19	59		20	44		15	42
	20	49		20	45		15	42
	20	49		20	45		15	42
	20	50		20	45		15	44
	20	51		20	46		15	44
	20	51		20	46		15	46
	20	53		20	47		16	37
	20	53		20	47		16	39
	20	53		20	48		16	40
	20	53		20	49		16	40
	20	53		21	41		16	40
	20	53		21	42		16	41
	20	53		21	42		16	41
	20	53		21	42		16	41
	20	53		21	43		16	42
	20	54		21	43		16	43
	20	54		21	44		16	43
	20	54		21	44		17	38
	20	55		21	44		17	38
	20	55		21	45		17	39
	20	56		21	45		17	39
	20	56		21	45		17	39
	20	59		21	45		17	39
	21	48		21	45		17	39
	21	49		21	46		17	40
	21	50		21	46		17	40
	21	50		21	46		17	40

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
er	21	50	U	21	47	u	17	40
	21	50		21	48		17	40
	21	52		21	48		17	40
	21	52		21	48		17	41
	21	52		21	48		17	41
	21	52		21	49		17	41
	21	53		21	49		17	41
	21	53		21	49		17	42
	21	53		22	40		17	42
	21	53		22	41		17	42
	21	54		22	42		17	42
	21	54		22	42		17	42
	21	54		22	43		17	42
	21	55		22	44		17	42
	21	55		22	44		17	43
	21	55		22	45		17	43
	21	57		22	46		17	43
	21	57		22	46		17	44
	21	57		22	46		17	44
	22	48		22	46		17	45
	22	49		22	46		17	45
	22	50		22	46		18	35
	22	52		22	47		18	37
	22	52		22	47		18	38
	22	52		22	47		18	41
	22	53		22	47		18	41
	22	53		22	47		18	41
	22	53		22	48		18	42
	22	54		22	48		18	42
	22	54		22	49		18	42
	22	54		22	50		18	42
	22	55		22	51		18	42
	22	55		23	43		18	43
	23	48		23	43		18	43
	23	49		23	44		18	43
	23	50		23	44		18	44
	23	51		23	45		18	44
	23	51		23	45		18	45
	23	51		23	45		18	48
	23	51		23	46		19	39
	23	53		23	46		19	40
	23	53		23	46		19	40
	23	53		23	46		19	40
	23	53		23	47		19	40
	23	54		23	47		19	41
	23	55		23	49		19	41
	23	55		23	49		19	43
	23	55		23	52		19	43
	23	55		24	43		19	44

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
er	23	57	U	24	44	u	19	44
	23	57		24	46		19	45
	23	58		24	46		20	38
	26	51		24	48		20	40
	27	52		25	46		21	41
				25	49		21	41
				26	47		22	40
				26	47		22	41
				26	47			
				27	49			
uh	22	45	aw	24	37	a	22	37
	23	43		24	40		22	38
	23	44		24	41		23	35
	23	46		24	43		23	36
	23	47		24	43		23	36
	23	47		24	44		23	36
	23	48		24	45		23	38
	23	50		24	46		23	38
	23	50		24	48		23	40
	23	51		25	39		23	41
	24	39		25	39		24	34
	24	42		25	40		24	34
	24	43		25	41		24	35
	24	45		25	41		24	35
	24	46		25	41		24	35
	24	46		25	41		24	36
	24	46		25	42		24	36
	24	47		25	42		24	36
	24	47		25	42		24	36
	24	48		25	42		24	37
	24	49		25	42		24	37
	24	49		25	42		24	37
	24	49		25	42		24	37
	24	50		25	43		24	38
	24	50		25	43		24	39
	24	51		25	43		24	40
	24	55		25	43		24	41
	25	40		25	43		25	34
	25	43		25	43		25	35
	25	44		25	44		25	36
	25	44		25	44		25	37
	25	45		25	44		25	37
	25	45		25	44		25	38
	25	46		25	45		25	39
	25	46		25	45		25	39
	25	46		25	46		25	40
	25	47		25	46		25	43
	25	47		25	46		25	43

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
uh	25	47	aw	25	47	a	26	33
	25	49		26	40		26	34
	25	51		26	40		26	35
	26	39		26	40		26	37
	26	41		26	41		26	37
	26	41		26	41		26	38
	26	42		26	41		26	38
	26	43		26	41		26	38
	26	43		26	42		26	38
	26	44		26	42		26	39
	26	44		26	42		26	39
	26	45		26	42		26	40
	26	45		26	42		26	40
	26	45		26	42		26	40
	26	47		26	43		26	40
	26	47		26	43		26	41
	26	47		26	43		26	41
	26	47		26	43		26	41
	26	47		26	43		26	43
	26	48		26	43		26	43
	26	48		26	44		26	44
	26	48		26	44		27	32
	26	48		26	44		27	32
	26	49		26	44		27	33
	26	49		26	44		27	34
	26	49		26	44		27	34
	26	49		26	44		27	35
	26	49		26	44		27	35
	26	50		26	44		27	35
	26	50		26	45		27	35
	26	50		26	45		27	36
	26	51		26	45		27	36
	26	52		26	45		27	37
	26	64		26	47		27	37
	27	41		26	47		27	37
	27	43		26	49		27	37
	27	47		27	39		27	37
	27	47		27	40		27	38
	27	47		27	41		27	38
	27	47		27	41		27	39
	27	48		27	43		27	41
	27	49		27	43		27	41
	27	49		27	43		28	34
	27	50		27	44		28	34
	28	45		27	44		28	34
	28	45		27	48		28	35
	28	46		27	48		28	36
	28	47		27	48		28	36

Phoneme	F1	F2	Phoneme	F1	F2	Phoneme	F1	F2
uh	28	47	aw	27	50	a	28	36
	28	49		28	39		28	36
	28	51		28	40		28	37
	28	54		28	40		28	38
	29	45		28	42		28	40
	29	49		28	43		29	32
	29	49		28	44		29	34
	29	50		28	46		29	38
	29	52		29	47		29	39
				29	49		29	42
				30	48		30	42