1989

Automatic formant labeling in continuous speech

Elizabeth A. Richards

Automatic Formant Labeling in Continuous Speech

by

Elizabeth A. Richards

Submitted to the Graduate Department of Computer Science of the Rochester Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science.

Approved by:

Dr. James M. Hillenbrand

John A. Biles

Dr. Peter G. Anderson

Title of Thesis: Automatic Formant Labeling in Continuous Speech

I, Elizabeth A. Richards _____ hereby grant my permission to the Wallace Memorial Library of R.I.T to reproduce my thesis in whole or in part. Any reproduction will not be used for commericial use or profit.

Date 9/22/89

Abstract

This thesis was developed out of a need to reduce the time required to correct Linear Predictive Code (LPC) data used for training a formant tracker. A program was written to select peaks from LPC data and interpret them as F1, F2, and F3, using knowledge about the phonetic transcription, the sex of the speaker, knowledge about individual phonemes, and a few heuristics. The system was tested on a database of eight speakers, four male and four female, each of whom produced ten sentences. This data set comprised 1,011 resonant phonemes covering 17,363 5-msec. frames. Overall the system correctly matched F1 in 98.9% of the frames, F2 in 92.2% of the frames, and F3 in 88.8% of the frames.

Table of Contents

CHAPTER 1

THESIS DESCRIPTION

The larynx, the source of speech sounds, creates an acoustic wave consisting of energy at many different frequencies. The speech organs of the mouth, moving to articulate individual vowels and consonants, act like a filter by suppressing sound energy at some frequencies while letting energy at other frequencies through. Peaks in the vocal tract frequency response curve are called resonances or formants. In most cases vowels can be identified by the frequency of the first two formants, but in a few instances the third formant is also necessary [HILL87]. Formant tracking, therefore, can provide a powerful source of information to a speech understanding system. If formants can be accurately tracked, then vowels should be able to be correctly labeled.

The formant tracker that is part of the Speech Understanding System being developed at the RIT Research Corporation requires training and later testing on valid formant frequency data [GAYV87, GAYV88]. The data, peaks and their corresponding amplitudes which have been extracted from the Linear Predictive Code Spectrum, have two major flaws: spurious peaks and missing formants. At the present time, these files must be edited by an expert, who has the following information about the utterance: the spectrogram, a plot of the LPC peaks, the label file which identifies each phoneme with its starting and ending times, knowledge about the expected frequency values for the first three formants of each phoneme, and information about how specific phonemes may affect other phonemes. After

displaying the spectrogram, the plot of the LPC peaks, and the labels, the expert would locate any flows of continuity for F1, F2, and F3 over several phonemes. He would then look at specific phonemes to determine whether the frequency values might be valid based on his knowledge of estimated frequency values for that phoneme and the coarticulatory effects of the surrounding phonemes. Any peaks whose frequency values seemed correct for the formant in question would be traced and written to a file.

Correcting the data files is time consuming and costly. This thesis was developed out of a need to reduce the cost and the human expert's time required to edit the database used to train and test the formant tracker. The purpose of this project is to decide which peaks from Linear Prediction Coding (LPC) Spectrum are the first three formants. The system has knowledge of the sex of the speaker and the phonetic transcription of each utterance which includes the starting time for each phoneme, along with information about individual phonemes and how they affect adjacent phonemes.

If an expert can make decisions as to which data points are actually formants and which are not, it should be possible to develop software to perform a similar task, provided that the computer has the same knowledge the expert has at his disposal. The computer can store information on each phoneme; it can extract the adjacent phonemes and retrieve information on how these phonemes may affect the target phoneme. The advantages the human expert has over an expert system are the visual cues from the spectrogram and the plot of the LPC data points. The expert system must compensate in some way, such as by applying heuristics.

Chapter 2 contains background information about the production and categorization of vowels, the effects of coarticulation in general, the specific effects of

coarticulation in diphthongs, a general discussion of the problems related to semivowels and nasals, and information about formant estimation. In chapter 3 the implementation of the system is discussed. Chapter 4 presents the results of the current system. Chapter 5 suggests some possible improvements that could be made to the system and discusses some insights gained from this work.

CHAPTER 2

BACKGROUND

2.1 Vowels

From a physiologic point of view the production of speech is governed by three major components: 1) the subglottal or respiratory system, 2) the larynx, and 3) the supralaryngeal vocal tract (see Figure 2-1).

The respiratory system can be compared to a balloon. As a balloon takes in air, the skin of the balloon expands. Likewise, as air is taken into the lungs, the walls of the lungs expand. When a balloon deflates, the stored-up energy is released, and air is forced out because of the elasticity of the balloon. The lungs react in a similar fashion when exhalation occurs: a steady stream of air is forced out. This air flow is the power source in the production of speech [LIEB77].

The steady current of air from the lungs is forced through the larynx, which contains the vocal folds. The vocal folds are closed until the air pressure from the lungs builds up and forces them open. After air has escaped, pressure from the lungs is reduced and the force of the laryngeal muscles closes the glottis. During speech there is a constant opening and closing of the glottis, converting a relatively steady current of air into a series of puffs of air.

The airflow then travels through the supralaryngeal vocal tract, which acts like an acoustical filter modifying the sound produced at the larynx. The size and shape of the vocal tract determine what the filtering properties will be.

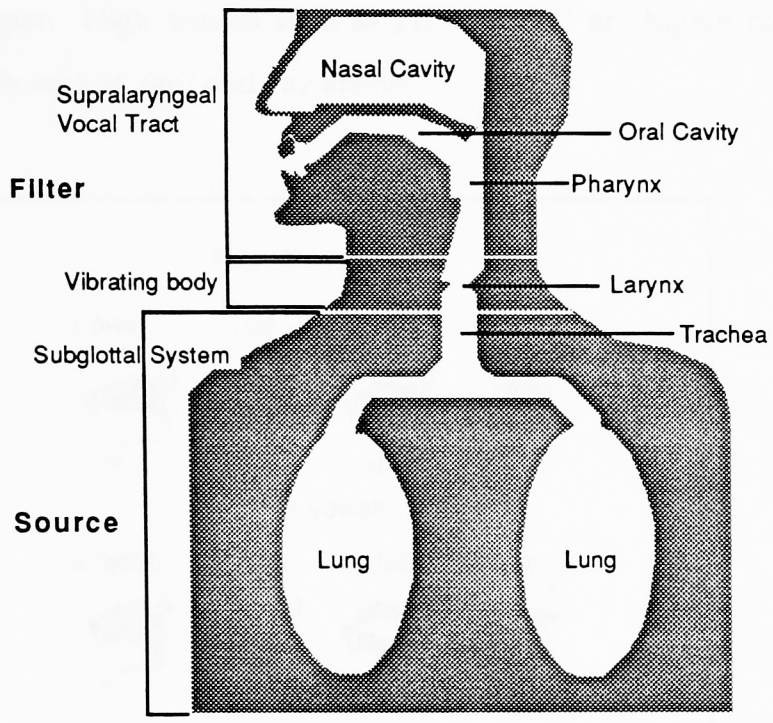


Figure 2 - 1
The Three Physiologic Components of Human Speech Production [LIEB77]

Vowels are produced by acoustically energizing the vocal tract at the level of the larynx while keeping the vocal tract relatively open. Consonant sounds, on the other hand, are produced with the vocal tract partially or completely occluded. The articulation of vowels is classified according to the major point of constriction in regards to position (front, central, or back) and height (high, mid, or low) of the tongue. As seen in Figure 2 - 2, the vowels /i, I, e, ɛ, ae, a/ are produced with the point of maximum tongue constriction near the alveolar ridge, and are, therefore,

called front vowels.[1] The vowels /u, U, o, ɔ, a/ are produced with the point of maximum tongue constriction near the velum and are called back vowels. The remaining vowel sounds, /ʌ, ɜ, ə, ɚ/, are produced in the area of the hard palate. These are called the central vowels. The degree of constriction is determined by the height of the tongue. High vowels such as /i/ and /u/ are highly constricted, whereas low vowels such as /ae/ and /a/ are not.

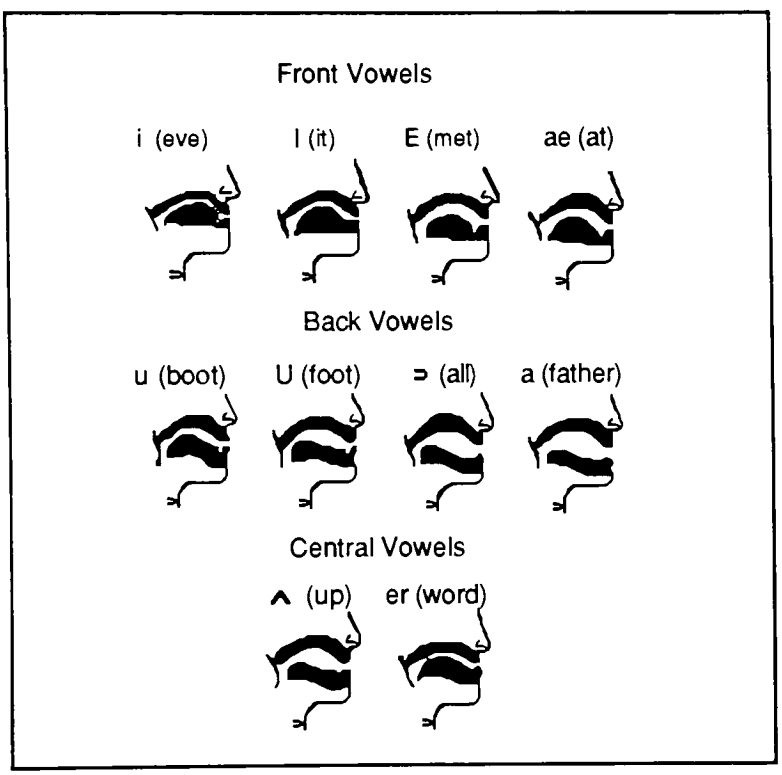


Figure 2 - 2
Schematic Vocal Tract Profiles for the Production of English Vowels [FLAN72]

---

[1]See Appendix A for phonemes represented in C-MU notation and the International Phonetic Alphabet.

In addition, the relative roundness (rounded or unrounded) of the lips and the tension of the tongue muscles (tense or lax) contribute to the characteristics of each vowel. Vowels that are rounded (e.g., /u, U, o/) are not only constricted in the mouth, but also at the lips. Tense vowels are longer in duration and carry more sound energy than lax vowels. In English most tense vowels have a corresponding lax vowel, that is, the description of these pairs of vowels differ only in the tense/lax category. For example the vowel /i/ is described as high, front, unrounded, and tense, whereas the vowel /I/ is described as high, front, unrounded, and lax.

Source-Filter Theory relates the physiologic description of speech production to the acoustic sound wave that actually creates speech. It views the vocal tract as a tube, closed at one end with a sound source (the larynx) and open at the opposite end (the lips). In the physiologic model the larynx changes a steady stream of air into a series of puffs. This model considers the larynx to be the excitation source. The acoustical wave is set up by the rapid changes in the physiological volume velocity in the glottis. At the source the acoustic volume velocity and the physiological volume velocity are proportional. The acoustical wave, which radiates at a velocity of about 34,000 cm/sec, instantaneously overtakes the physiological wave, which radiates at a velocity of about 5 cm/sec. According to Minifie [MINI73] this concept can be compared to a swimmer diving into a pool. When he enters the water, he disturbs the medium and sets up a wave in the water. His entry into the water also sets up an acoustic wave which will travel either in the water or in the air. If a person is standing at the other end of the pool from the diver, he will hear the acoustical wave (the sound) long before the water wave reaches him. The same driving force sets up two types of waves within a medium. These waves travel with different velocities, depending upon the nature of the energy in the wave [MINI73].

It is the acoustical volume velocity that creates the resonances in the vocal tract for speech production.

The acoustic wave then travels through the vocal tract, which acts as a filter. The speech organs of the mouth move to articulate individual vowels by changing the size and shape of the vocal tract. This physiologic change suppresses sound energy at some frequencies while allowing energy at other frequencies to pass through.

Vowels are traditionally defined by formant frequency patterns. Formant frequencies are the natural frequencies of the decaying sinusoidal responses of a resonance of the vocal tract. Each one of the formant frequencies is specified by a number, such as F1, F2, or F3. Discrimination of most vowels and some consonants is accomplished largely through the specification of the first two formants. However, in some cases additional information is provided by the third formant [HILL87]. The formant frequencies are related to the length of the vocal tract because the shape and length of the supralaryngeal passages act as an acoustic filter. A longer tract, as is common with men, is associated with lower frequencies. A shorter tract, more common in children and to some degree women, is associated with higher frequencies.

One of the most thorough vowel studies using the sound spectrograph and one which is now considered classic was conducted by Peterson and Barney [PETE52]. Peterson and Barney studied vowels in an /h V d/ (h vowel d) context to investigate the relationship between the phoneme intended by the speaker and that identified by the listener, and to relate these phonemes to acoustical measurements of 10 American English vowels produced by 76 speakers. These speakers, including 33 men, 28 women, and 15 children, recorded two repetitions each of ten vowels (76 x

10 x 2) for a total of 1520 words. Words were randomized and presented to a group of 70 listeners who were asked to classify the words into one of the ten categories. In addition, the words were analyzed by means of the sound spectrograph.

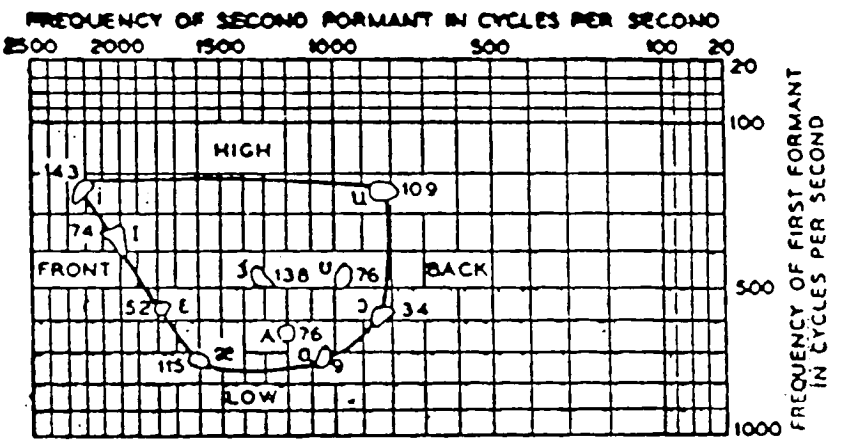


Figure 2-3
Vowel Loop with Numbers of Sounds Unanimously Classified by Listeners [PETE52]

Peterson and Barney measured the fundamental frequency, the formant frequencies of F1 - F3, and relative formant amplitudes of each utterance. Figure 2-4 shows a plot of F1 and F2 for one calling of all the words for each speaker. The / ɝ / sound can be distinguished from the other vowels if the third formant is used. The distribution of points in figure 2-4 is continuous. There is no definitive break between each of the vowels. Indeed, even disregarding / ɝ /, there is considerable overlap for some vowels (e.g. /a/ and / ɔ /).

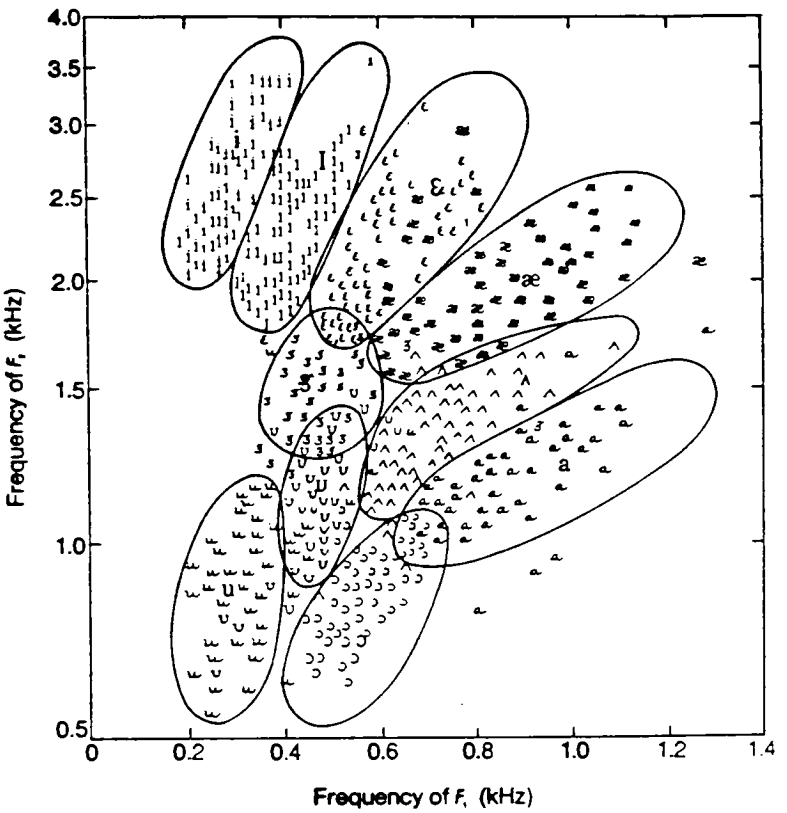


Figure 2-4
Frequency of Second Formant Versus Frequency of First Formant for 10 Vowels by 76 Speakers [PETE52].

2.2 COARTICULATION

Coarticulation is the process by which the articulatory characteristics of one sound are modified by adjacent sounds. It is assumed that when a phoneme is produced in isolation, it will reach an articulatory target, but if the targets of two adjacent phonemes are far apart, neither target location is apt to be reached. This response is called undershoot [STEV66].

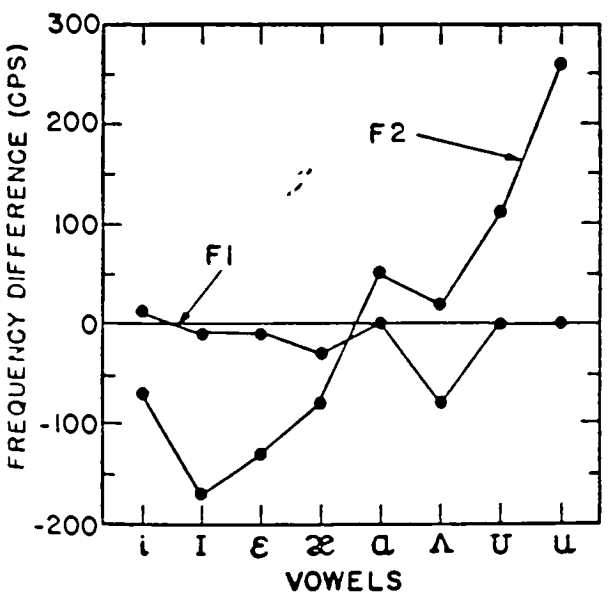


Figure 2-5
Differences between Average Target Values for F1 and F2 of Eight Vowels in Null Contexts and Average Values for the Same Vowels in 14 Consonantal Contexts [STEV63]

The extent to which the time-varying configuration falls short of the target vowel configuration will be greater when the duration of the vowel is less and will increase with the effective distance that must be traversed between the vowel and

consonant configurations in the syllable [STEV63]. For long vowels like /i/, /ae/, and/a/ the undershoot in F2 is less than for the short vowels / I /, / E /, / ^ /, and / U /. The effect is demonstrated in Figure 2-5, where zero represents the target value and the points for each vowel indicate the distance (positive or negative) F1 and F2 missed that target value. The result of undershoot is a phenomenon called centralization. The values that are reached are in a more central part of the vowel chart because of the coarticulatory effects of adjacent phonemes [STEV63].

Coarticulation works in both directions. The influence of one sound on the preceding sound is called anticipatory or forward coarticulation. For example, in the word 'tenth' the /n/ becomes dentalized because of the dental features of 'th.' Retentive or backward coarticulation is the influence of one sound on the production of the succeeding sound. In the word 'hypnotize' the /n/ often becomes an /m/ because of the influence of the labial /p/.

Öhman supported this concept in his study about the effects of coarticulation on stop consonants in a vowel-consonant-vowel (VCV) environment. In English he studied the vowels /i/, /a/, and /u/ in the context of /d/ and /g/. He concluded that the terminal frequencies of the formants in VCV utterances depended not only on the consonants, but on the entire vowel context. He postulated that the production of the consonant involves concomitant articulatory adjustment partially anticipating the configuration of the succeeding vowel, and, therefore, that the articulation of the consonant and the following vowel must be active simultaneously. The medial consonant configuration is also anticipated during the initial vowel. He also observed this anticipatory effect at the beginnings of prolonged initial vowels, therefore verifying the effect of backward coarticulation.

The effects of consonantal context on the vowel formant frequencies were studied by Stevens and House [STEV63]. They conducted a study similar to that of Peterson and Barney [PETE52], but instead of measuring formant frequencies at centrally located points in the /h_d/ context, they studied vowels in 14 different consonantal contexts produced by three speakers. They found that the consonantal context causes systematic shifts in the vowel formant frequencies depending upon the place of articulation of the consonant, its manner of articulation, and its voicing characteristic.

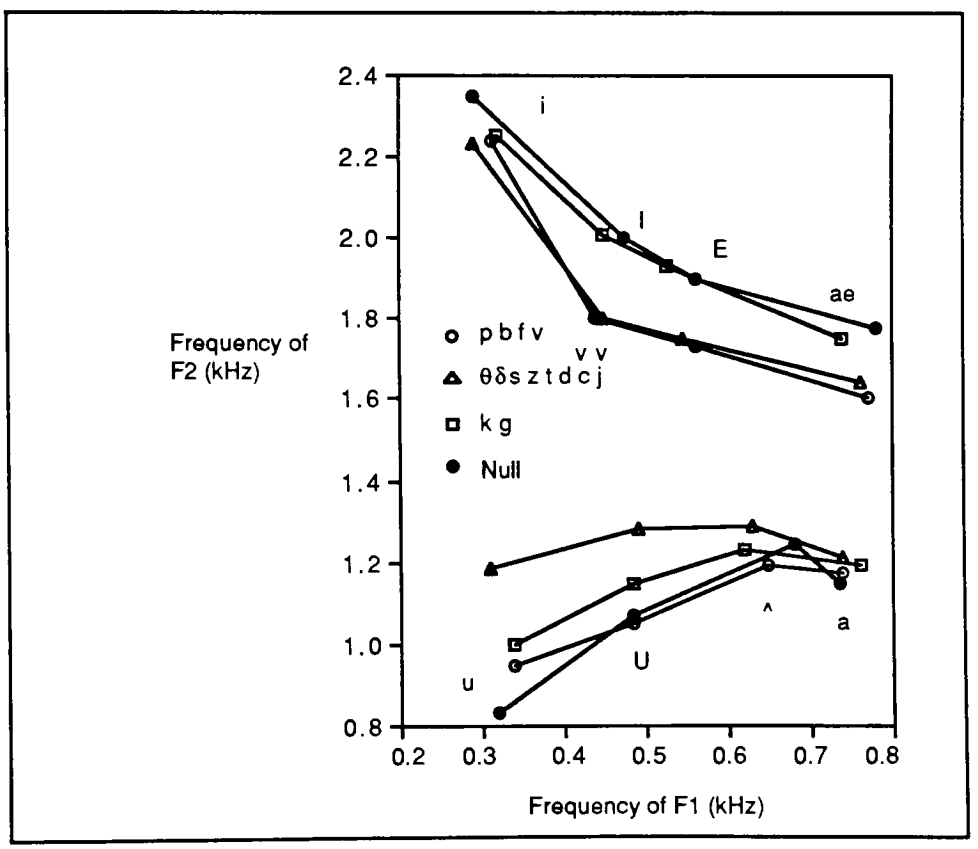


Figure 2-6
Values for 8 Vowels Plotted to Demonstrate the Effect of Place of Articulation of the Consonantal Context [STEV63]

In Figure 2-6 [STEV63] the values for F1 and F2 are plotted to demonstrate the effect of place of articulation of the consonantal context. The contexts are specified according to place of articulation: velars, postdentals, and labials. Values for vowels in the null environment (/h_d/) are also included. The values for the first formant are not generally affected by the place of articulation. However, the second formant is not as stable. The amount of deviation from the null environment is dependent on the individual vowel. The phonemes /u/ and /U/ exhibit the greatest variation in F2, whereas /a/ and /i/ show little effect from the consonantal context.

In Figure 2-7 [STEV63] the formant frequencies are examined according to the voicing characteristic and the manner of production of the consonantal context. The results demonstrate the displacement of F2 toward a central position when average data for vowels in consonantal contexts are compared with data for vowels in null contexts.

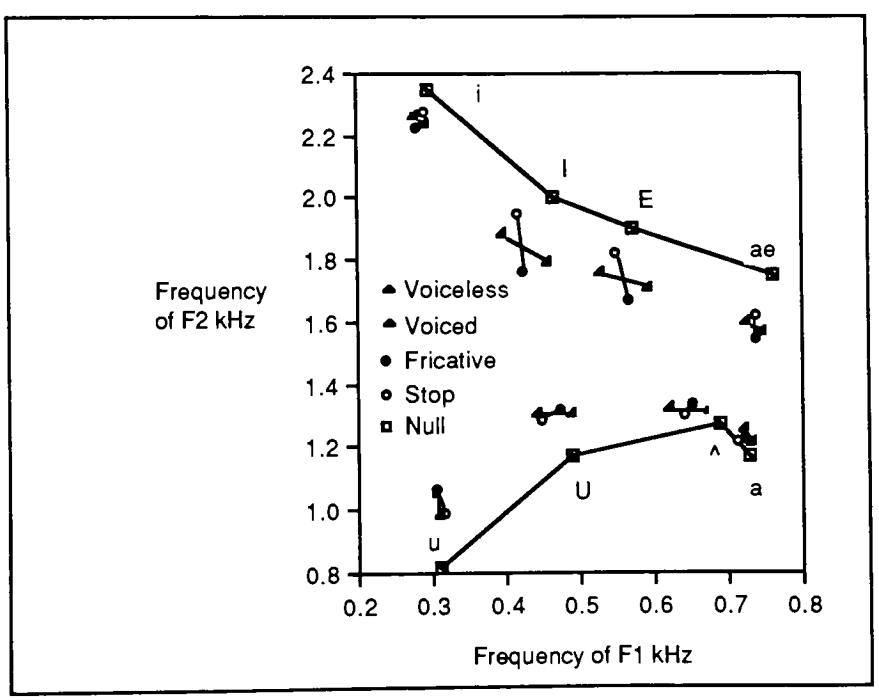


Figure 2-7
Values for F1 and F2 for 8 Vowels Plotted to Demonstrate the Manner of Production and Voicing Characteristics of the Consonantal Context [STEV63]

Their data showed that formant frequencies are modified in a systematic way by the consonantal context. Therefore, Stevens and House theorize coarticulation should be predictable by phonological rules that show how a phonemic sequence will be realized by allophones. An allophone is a variation of a phoneme which is often caused by coarticulation.

Most studies of coarticulation focus on phonemes in an isolated VCV or CVC environment. Considerably more difficult is the problem of coarticulation in continuous speech. The timing of the articulators can be more carefully planned in the CVC (or VCV) case than in continuous speech. That means vowels can be longer in CVC's. Also, consonant clusters in continuous speech can have a cumulative effect on vowel formants. Prepausal lengthening and duration changes caused by stress mean that the vowel durations are much more variable than in the isolated CVC word case [HIER86].

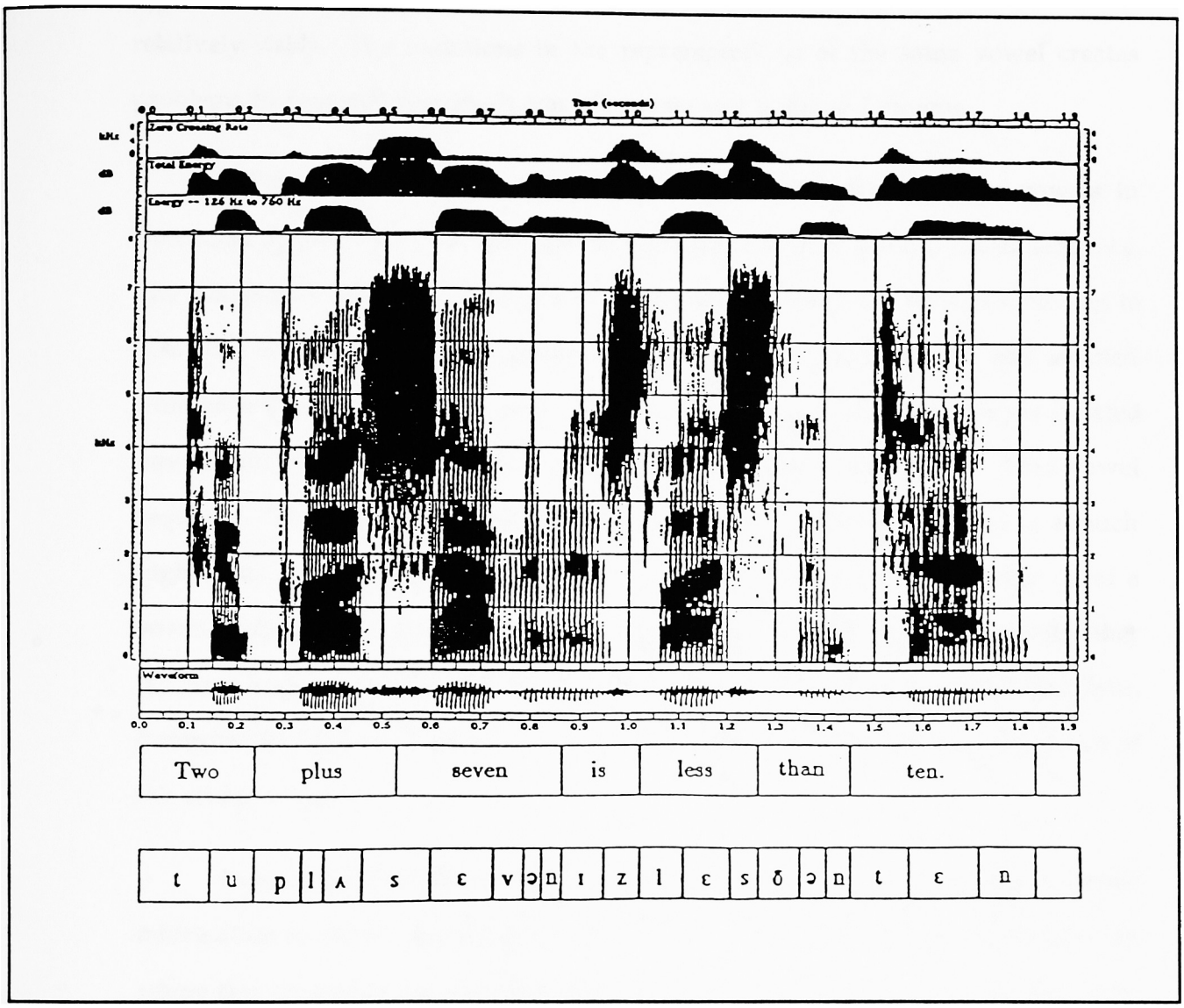


Figure 2-8
Spectrogram of the Sentence "Two plus seven is less than ten." which Illustrates Variations in the Representation of the Phoneme / ɛ / [ZUE85]

As can be seen in Figure 2-8 [ZUE85] the phoneme / ɛ / appears in three places and in each place it has different characteristics. In the context of /s/ and /v/ F1 is rising while F2, F3, and F4 are falling. In the context of /l/ and /s/ F1, F2, and F4 are rising while F3 is stable. In the context of /t/ and /n/ all four formants are

relatively stable. The variations in the representations of the same vowel creates problems in recognizing vowels from formants and tracking formants.

Kuwabara [KUWA85] performed an experiment with Japanese vowels in connected speech. He recorded speech units consisting of a three-vowel sequence. The continuous three-vowel sequences were separated from the spoken sentences to form the stimuli for his perceptual experiment. The middle vowel was isolated from each three-vowel speech unit. The listeners were asked to identify the isolated vowel and then were asked to identify the middle vowel in the three-vowel sequence. The middle vowel in the three-vowel sequence was identified at a much higher rate than the isolated vowel. The result indicates that the perception of a vowel in connected speech is highly context dependent. This finding indicates that the time-varying patterns of acoustic features such as formant trajectories from immediately adjacent vowels may have a lot to do with the correct identification of the center vowel.

In his paper Kuwabara describes an approach for incorporating the contextual information to reduce the ambiguity of vowels. Figure 2-9 is a plot of the vowels where this technique has not been applied. As can be seen, the distribution of the vowels in the F1-F2 space in Figure 2-9 is much broader than that in Figure 2-10 which shows a plot the vowels where his technique has been applied.

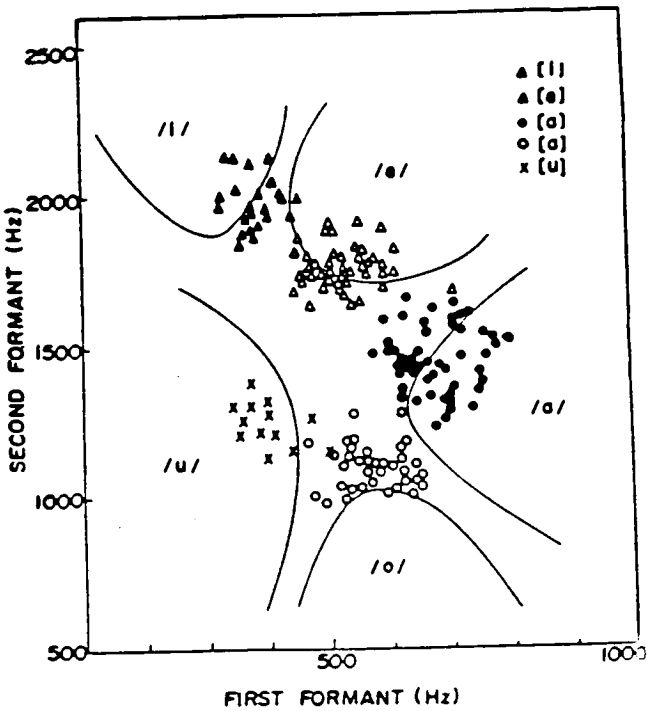


Figure 2-9
F1 F2 Formant Frequencies for Japanese Vowels [KUWA85]

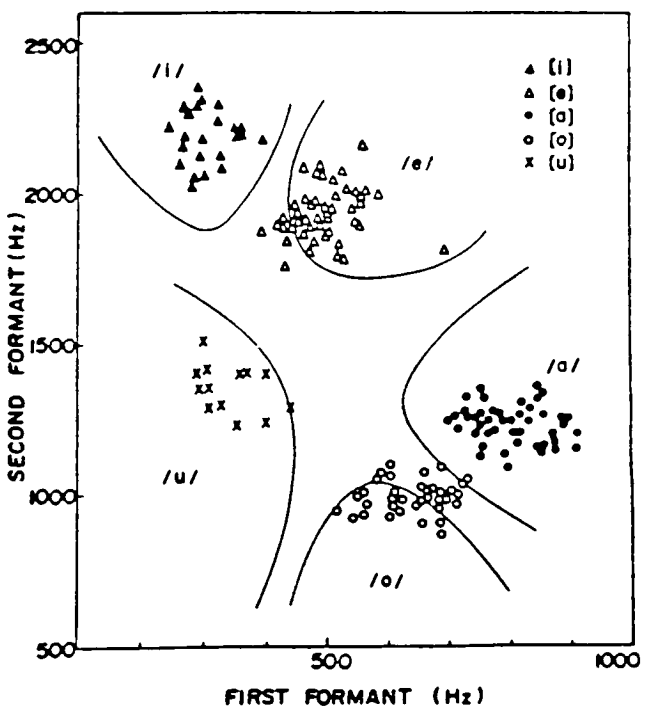


Figure 2-10
Modified Formant Frequencies of the same Vowels as in Figure 2 - 9 [KUWA85].

Because Kuwabara had such good results with Japanese vowels, Hieronymus and Majurski [HIER86] applied his method to American English. The results were not as successful. Hieronymus and Majurski attribute this to the simpler CVCV structure of Japanese, as opposed to American English which is more complicated. Japanese also has a much smaller vowel system.

2.3 DIPHTHONGS

A diphthong is the combination of two vowel sounds that is formed by moving the tongue from one position to another. The diphthongs in English are listed on the chart below using the International Phonetic Alphabet. (IPA) The International Phonetic Alphabet is used in most of the diagrams. [2]

International Phonetic Alphabet	Sample Word
/au/	bought
/ei/	bait
/ai/	bide
/ɔi/	boy
/ou/	boat
/ju/	you

Diphthong Representations
Table 2-1

The phoneme /ju/ is not actually a diphthong because it consists of a semivowel followed by a vowel rather than two vowels. However, because the coarticulatory effects for /ju/ are very similar to those for diphthongs, this syllable will be treated as a diphthong in the present study On the vowel chart these sounds do not appear as points but as trajectories from an initial position to a final position

[2]A complete table of phoneme notations can be found in Appendix A.

(see Figure 2 - 12). The dynamic nature of diphthongs can be clearly seen in the spectrograms in Figure 2-11.

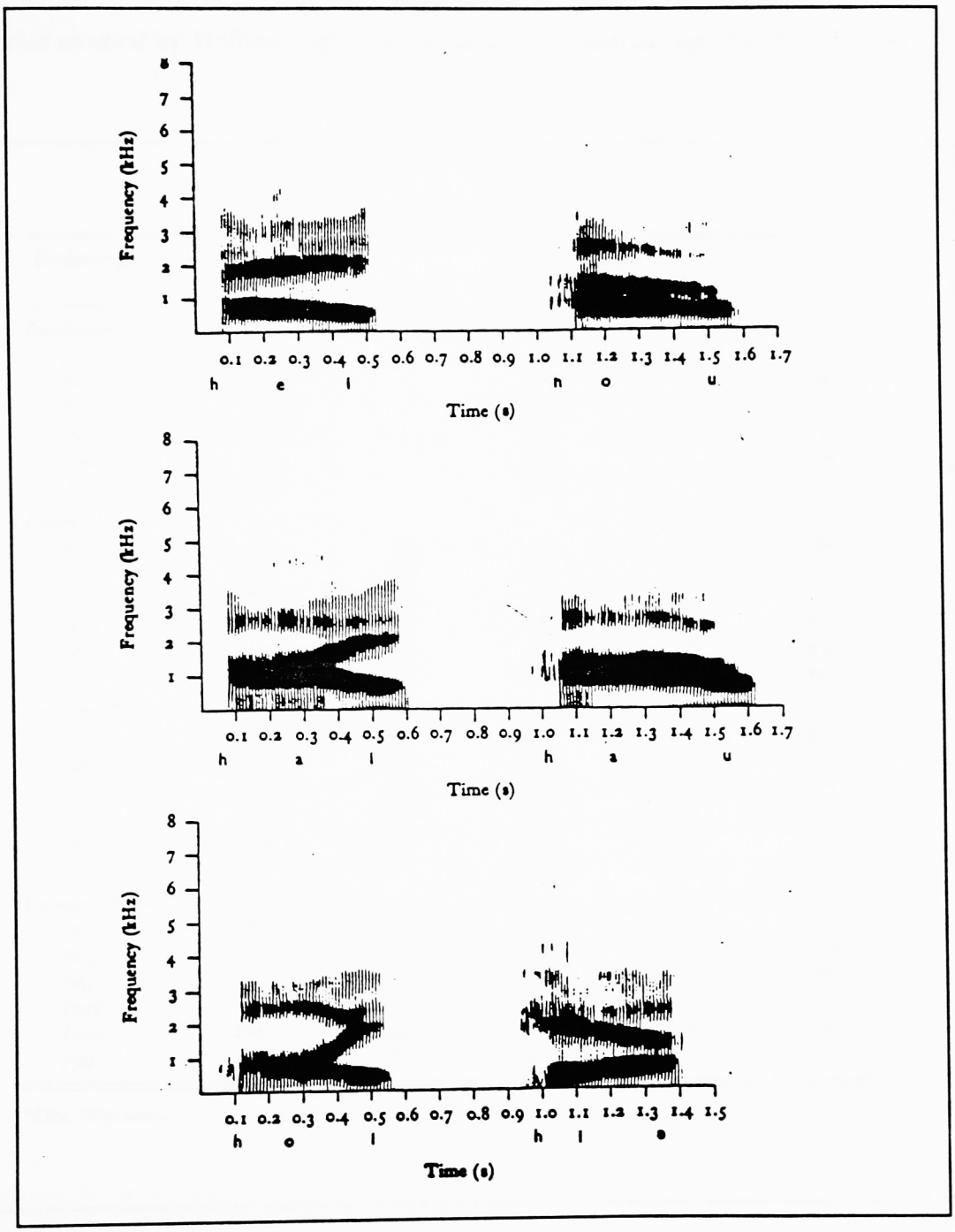


Figure 2-11
Spectrograms of English Diphthongs[FRY79].

The first target is usually longer than the second, but the transition between the targets is longer than either target position [LEHI61]. The temporal variations were also studied by Holbrook and Fairbanks and the results reported in Table 2-2.

Diphthong	*Sampling Point*				
	1	2	*3*	*4*	*5*
Fundamental					
/eɪ/	110		95		90
/aɪ/	115		98		98
/ɔɪ/	115		100		85
/oʊ/	118		102		95
/aʊ/	115		102		90
/ju/	118		108		92
Formant One					
/eɪ/	550	520	488	418	400
/aɪ/	750	710	735	682	572
/ɔɪ/	552	550	600	570	512
/oʊ/	565	535	505	472	465
/aʊ/	770	740	735	680	610
/ju/	330	342	338	352	400
Formant Two					
/eɪ/	2 032	2 078	2 088	2 125	2 228
/aɪ/	1 280	1 350	1 410	1 648	1 942
/ɔɪ/	835	888	1 062	1 558	1 908
/oʊ/	882	848	820	758	708
/aʊ/	1 400	1 320	1 210	1 062	888
/ju/	2 160	1 962	1 550	1 292	1 240*
Formant Three					
/eɪ/	2 650	2 660	2 650	2 662	2 710
/oɪ/	2 730	2 690	2 515	2 545	2 668
/ɔɪ/	2 525	2 410	2 350	2 290	2 492
/oʊ/	2 385	2 435	2 415	2 460	
/aʊ/	2 695	2 680	2 600	2 500	2 240
/ju/	2 568	2 390	2 275	2 285	

*One utterance.

Table 2 - 2

Median Frequencies of Diphthong Formants at Sampling Points [HOLB62].

Holbrook and Fairbanks [HOLB62] also studied the formant frequencies of diphthongs. As can be seen in Figure 2-12, there is extensive overlap of vowel areas.

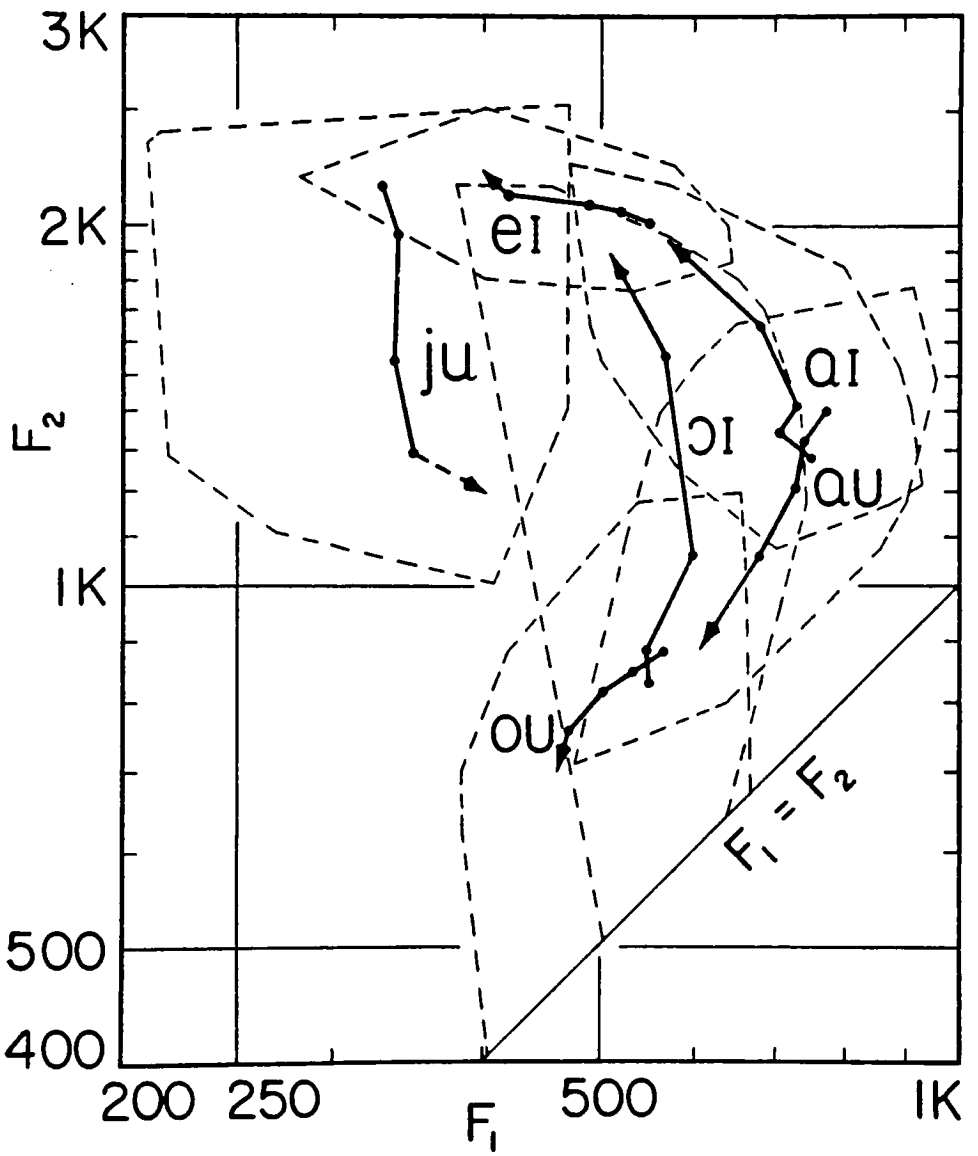


Figure 2-12

Central Regions of Variations of Frequencies of F1 and F2 in Diphthongs. [HOLB62]

Several interesting observations can be noted. The F1 range for /ai/ is similar to that of /au/, while the F2 ranges of /ai/ and /au/ overlap in part and then extend in opposite directions. In /ei/ and /ou/ the F2 ranges are widely separated, but they have the F1 range in common. In /ɔi/ and /ou/ F1 and F2 could not be distinguished as separate regions in some of the samples. In Figure 2-12 these values are represented by the points placed along the 45 degree line at the bottom of the Figure (F1 = F2) [HOLB62]. It was found in some samples of /ɔ/ only one formant could be identified in the lower frequency range [FAIR61]. The same one-formant characteristic was found by Holbrook and Fairbanks in six utterances of /ɔ i/ and in four of /ou/ [HOLB62]. The median values for the formant frequencies of the diphthongs have been connected and displayed against the outline of areas of the plot of the data points for F1 and F2.

Neither element comprising a diphthong is ordinarily phonetically identifiable with any stressed monophthong in English [LEHI61]. For example, the first element of /ai/ is neither an /a/ nor /ae/ and the second element is neither /i/ nor /I/. In Figure 2-13 the location of the formant frequencies for the diphthongs and the monophthongs are labeled.

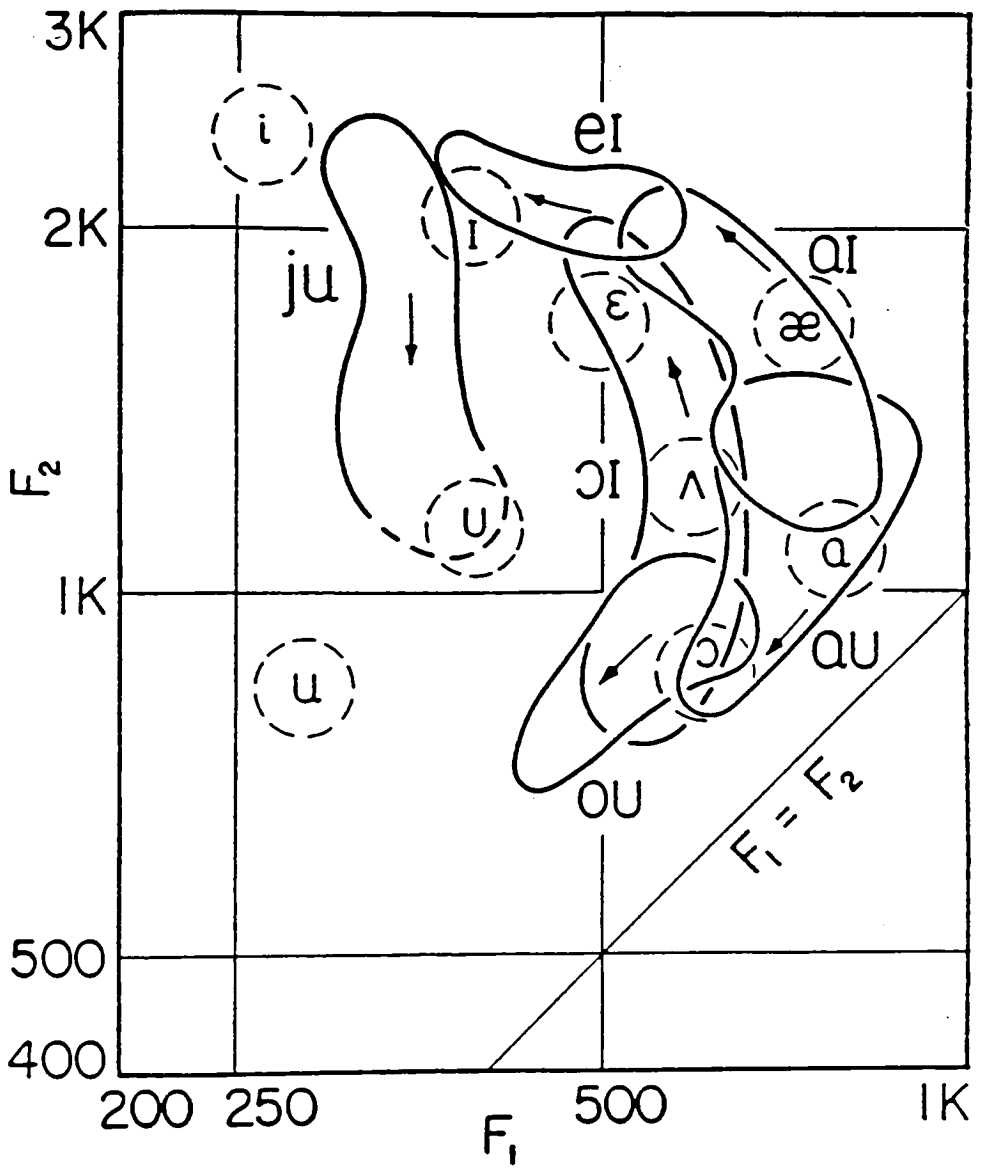


Figure 2-13
Location of Diphthongs in Relationship to Vowels. Dashed Circles Indicate Locations of Preferred Vowels. [HOLB62].

2.4 SEMIVOWELS

In the production of semivowels there is a greater degree of constriction than there is for vowels, but the constriction is not as great as it is for the other consonants. They are produced with vibration of the vocal folds and are, therefore, voiced sounds. They are in a fuzzy area between vowels and consonants, which creates several problems in trying to label them. Included in this group are /r, l, w, j/. The phoneme /j/ is sometimes represented as /y/. In both initial and final clusters these phonemes must occupy the position adjacent to the vowel (e.g. 'swat' or 'start'). The one exception is in words like 'world' and 'snarl' where /r/ comes between the vowel and /l/ [OCON57].

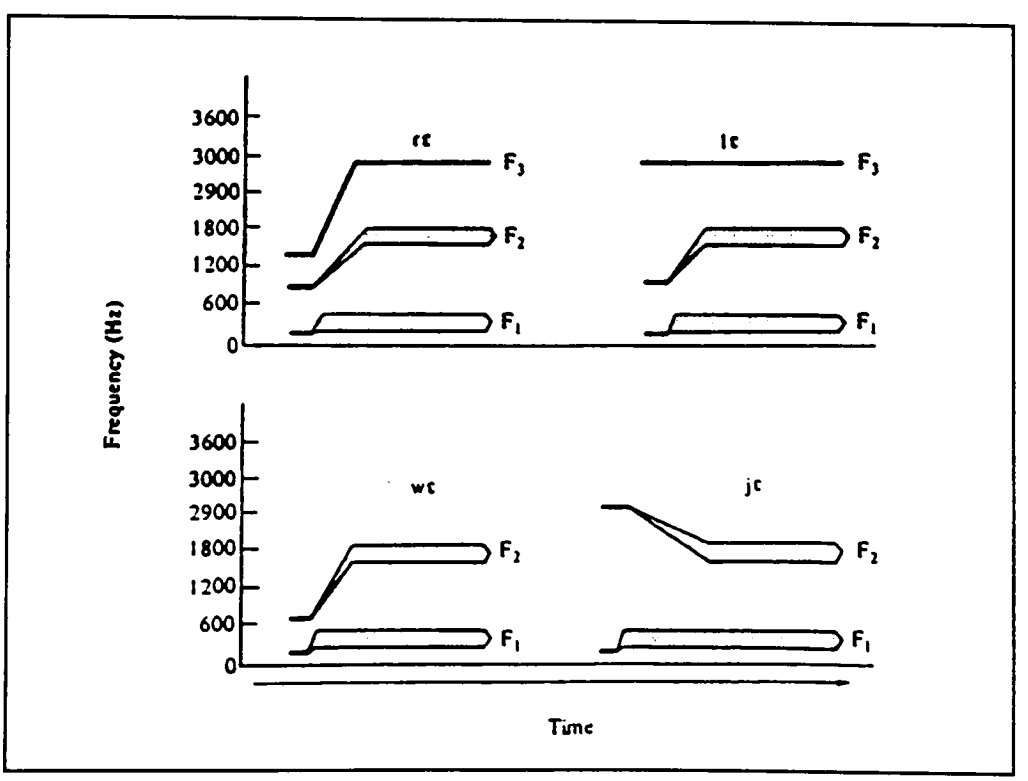


Figure 2-14
Simulated Spectra for the Semivowels /r/, /l/, /j/, and /w/ [OCON57]

Semivowels are primarily recognized by the transitional cues of the second and third formants, as can be seen in Figure 2-14. Semivowels are shorter in duration than vowels and therefore do not usually maintain a steady state. This makes it very difficult to locate them, especially if there are both forward and backward coarticulatory influences on them. Because formant transitions of semivowels are fairly smooth and much slower than those of other consonants, they exert a very strong influence on adjacent phonemes.

The formants for the phoneme /l/, also called a lateral, are dependent on vowel context. The second formant in particular follows F2 of the following vowel. When /l/ is in the final position, the second formant is still somewhat variable. The syllabic /l/ following consonants such as /k/, /g/, /p/, and /b/ appears more stable. In a comparison of the syllabic /l/ with the final /l/, the syllabic /l/ has the first formant in approximately the same frequency region as the final /l/. The second formant is definitely lower, and the third formant is usually slightly lower for the syllabic /l/ than for the final /l/ [LEHI 64].

The phoneme /r/ is also called a retroflex. One acoustic correlate of retroflexion is an exceedingly low third formant with a narrow separation between F2 and F3. Lehiste found this held true in her study for all allophones of /r/. The acoustic variations noted in the allophones of /r/ were correlated to the position of the sound sequence, such as initial, final, or syllabic. The initial allophones had lower formant frequencies for the first three formants. Although the initial allophones of /r/ do not seem to be influenced by the following vowel, the final allophones were quite dependent upon the preceding vowel [LEHI 64].

In her study Lehiste concluded that /w/ is similar to the vowel /u/, and that /j/ is similar to /i/. The first two formants were lower in frequency for the

semivowel than for the corresponding vowels. The third formant for /w/ was weaker or missing. For /j/ it was higher in frequency than for /i/. The /w/ and /j/ are more difficult to identify than their corresponding vowels because their duration is shorter and their formants are moving throughout most of the production of the sound. They also never occur in stressed positions as do their corresponding vowels [LEHI 64].

2.5 NASALS

The articulatory system for the production of nasals [m, n, ng], shown in Figure 2-15 [FUJI62], consists of three subsystems:

1) the pharynx extending from the glottis to the velum,

2) the oral cavity with a complete closure at the anterior end, and

3) the nasal cavity including the nasopharynx and nasal passages that are terminated by the radiation impedances of the nostrils.

The nasals are articulated by lowering the soft palate so that the air passage which leads out through the nasopharynx and nostrils is open. At the same time the oral cavity is completely occluded.

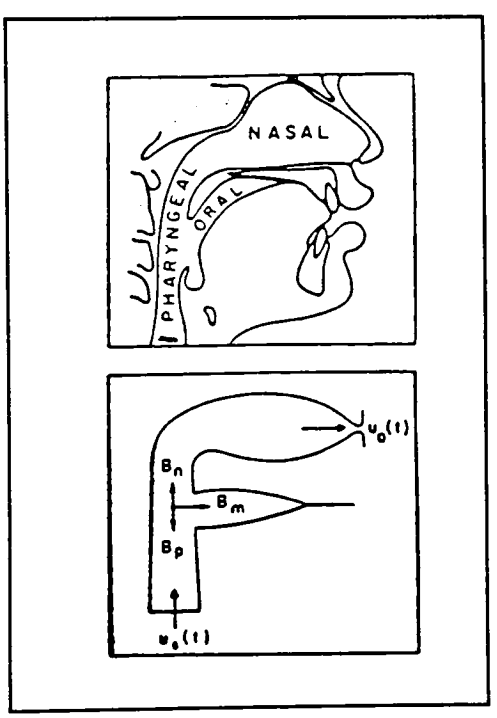


Figure 2-15
Articulatory System for the Production of Nasals [FUJI62]

The pharynx and the nasal cavity are the primary resonators for nasal consonants. The occluded oral cavity is a side branch. The length of the nasal cavity is longer than the oral cavity, and consequently the formants are generally lower. Because the nasal cavity has a fixed shape, and the pharynx does not alter appreciably for the different nasals, one would assume that the formant frequencies would be rather constant. However, the side cavity introduces an antiresonance, or zero. The location of the antiresonance varies depending on the place of articulation [HILL77] and the length of the nasal cavity [LASS82].

The soft linings of the nasal cavity cause increased damping of the acoustical signal, producing broader and more diffuse resonances. Fujimura [FUJI62] found the bandwidths for nasal murmurs comparable to or greater than those observed in vowels. Individual formants have different amounts of damping, and the bandwidths are not necessarily constant nor monotonically related to the frequency of the formant. Broader resonances are more difficult to find than narrow ones. [HILL – discussion 1/22/88]

Generally a nasal occurs adjacent to a vowel. These vowels will tend to be nasalized because the velum does not have enough time to open and close for each phoneme. For example, the word man is not said as [m ae n], but as [m $\tilde{\text{ae}}$ n] where the ~ indicates nasalization of the vowel. Oral vowels are represented by poles only. Nasals, on the other hand, are represented by poles and zeros. The complex relationship between poles and zeros creates unpredictable effects on nasalized vowels. For example, if an oral pole and a nasal zero coincide, they will cancel each other out.

Fujimura [FUJI62] characterized nasals as having the following three properties in common: 1) a low first formant around 300 Hz, 2) high damping of

the formants, and 3) high density of the formants in the frequency domain [FUJI62, p. 1874].

As a group nasals pose a major stumbling block with their unpredictability. As Fujimura summarizes

"the appearance of the spectra of nasal murmurs may vary considerably from one sample to another, depending on the individual nasal consonant and its context; the spectra also depend on the individual speaker who utters the sound, or even his temporary physiological state. The spectrum envelope can be altered significantly by a slight modification of the pole-zero pattern. The structure of the cluster, in particular, is of primary importance in determining the relative levels of the local spectral peaks in the middle-frequency range. The variability of the relative levels even within one nasal murmur presumably causes inherent difficulty for an automatic recognition scheme that is based on a straightforward analysis." [FUJI62, p. 1874].

2.6 FORMANT ESTIMATION

The formants, at least the first three, can represent a potentially powerful source of information for speech recognition. Information about the formants is contained in the spectral envelope. However, identifying the individual formants from the spectral envelope is a difficult task. Parsons [PARS86] described three problems associated with formant tracking:

1) Spurious peaks. Normally the maxima in the spectrum envelope are the actual formants. However, peaks that are not formants often exist.

2) Blends. Frequencies of adjacent formants may be too close to each other to resolve. In this case there is a deficiency of peaks.

3) High-pitched speech. High-pitched voices have relatively widely spaced harmonics and therefore provide fewer points from which to estimate the spectrum envelope [PARS86, p. 210].

There has been much research done in the area of estimation of formants from the spectral envelope, and the techniques vary greatly. An early formant extractor was based on estimation from filter-bank outputs [FLAN56]. The speech signal was filtered into multiple frequency bands (between 20 and 30), covering the frequency range of human speech. The output of each filter was a measure of the energy in that frequency band. One advantage of the filter-bank technique, and the reason it is still used today in speech recognition, is that filter banks can be designed to match the estimated frequency sensitivity of the ear more closely than linear prediction, a favored technique [PARS86].

Formant tracking has also been attempted using cepstrally smoothed spectra [SCHA70]. The cepstrum is a Fourier transform of a Fourier transform. Formants are estimated from the cepstrally smoothed spectra using frequency ranges for F1, F2,

and F3, amplitudes of the peaks, and the relationship of F1 and F2 amplitudes and F2 and F3 amplitudes. For the fundamental frequency and each of the first three formants a minimum (F#MN) and a maximum (F#MX) are set (see Figure 2-16). In addition, there is a variable for the amplitude of the fundamental frequency and each of the first three formants in the form of F#AMP. Formants are picked in order starting with F1. The search space for each of the formants is generally F#MN to F#MX unless the selected peak for the previous formant is greater than F#MN where # refers to the formant currently being searched for. Then the search space encompasses the current formant's search space in addition to the previous formant's search space. Because there is overlap between adjacent formants, this helps eliminate the problem of selecting the F2 peak as F1 and the F3 peak as F2. If F1 is greater than F2 or F2 is greater than F3, the respective values are reversed.

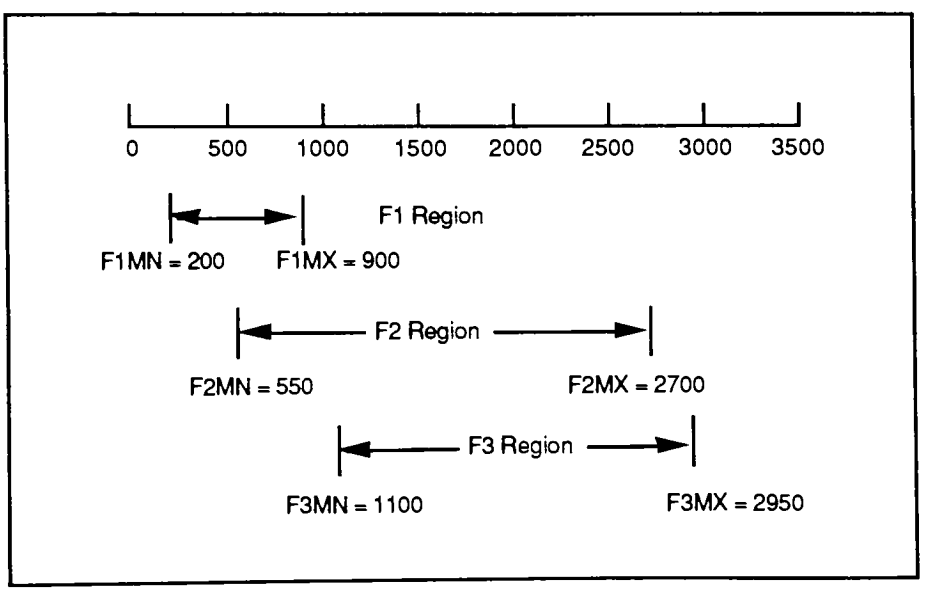


Figure 2-16
Frequency Ranges for the First Three Formants [SCHA70]

Estimation of F1: The peak with the highest amplitude in the frequency range 0 to F1MX is labeled as F0AMP. The difference between F0AMP and each peak in the F1 search space is computed. If that difference is less than 8.7dB, the peak is a possible candidate for F1. The peak among the candidate peaks with the highest amplitude is selected as F1. The amplitude of that peak becomes F1AMP. If no peak meets these criteria, the region from 0-900 Hz is expanded and enhanced using the Chirp Z-transform algorithm.

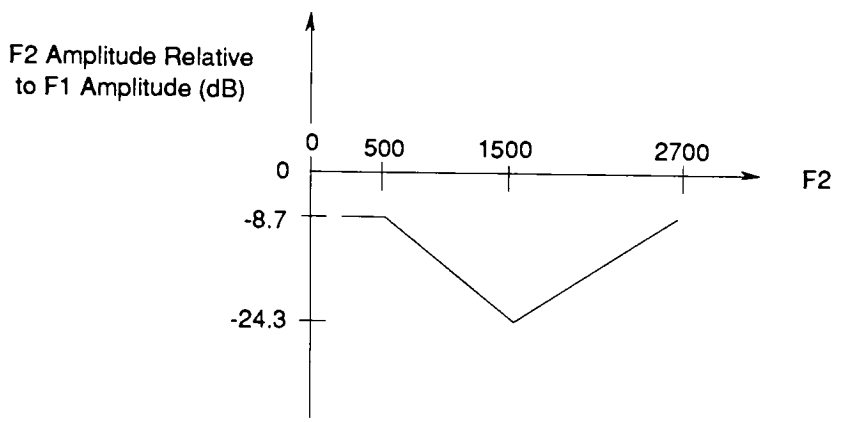


Figure 2-17
Threshold for Amplitude of F2 Peak Relative to the F1 Peak [SCHA70]

Estimation of F2: A peak is a candidate for F2 if it is within the F2 search space and the difference between F1AMP and its amplitude is within the threshold delineated by the frequency curve in Figure 2-17. The F2 candidate with the highest amplitude becomes F2 and its amplitude becomes F2AMP. If there are no candidate peaks, the Chirp Z-transform is performed.

Estimation of F3 : A peak is a candidate for F3 if it is within the F3 search space and the difference between F2AMP and its amplitude is within the -17.4dB if F2 is located

without the Chirp Z-transform. If F2 is located with the Chirp Z-transform, the threshold is set at -1,000 dB, essentially eliminating any threshold. The F3 candidate with the highest amplitude becomes F3. If there are no candidate peaks, the Chirp Z-transform is performed.

The formant and pitch period data obtained from this study were used to control a digital formant synthesizer [SCHA70].

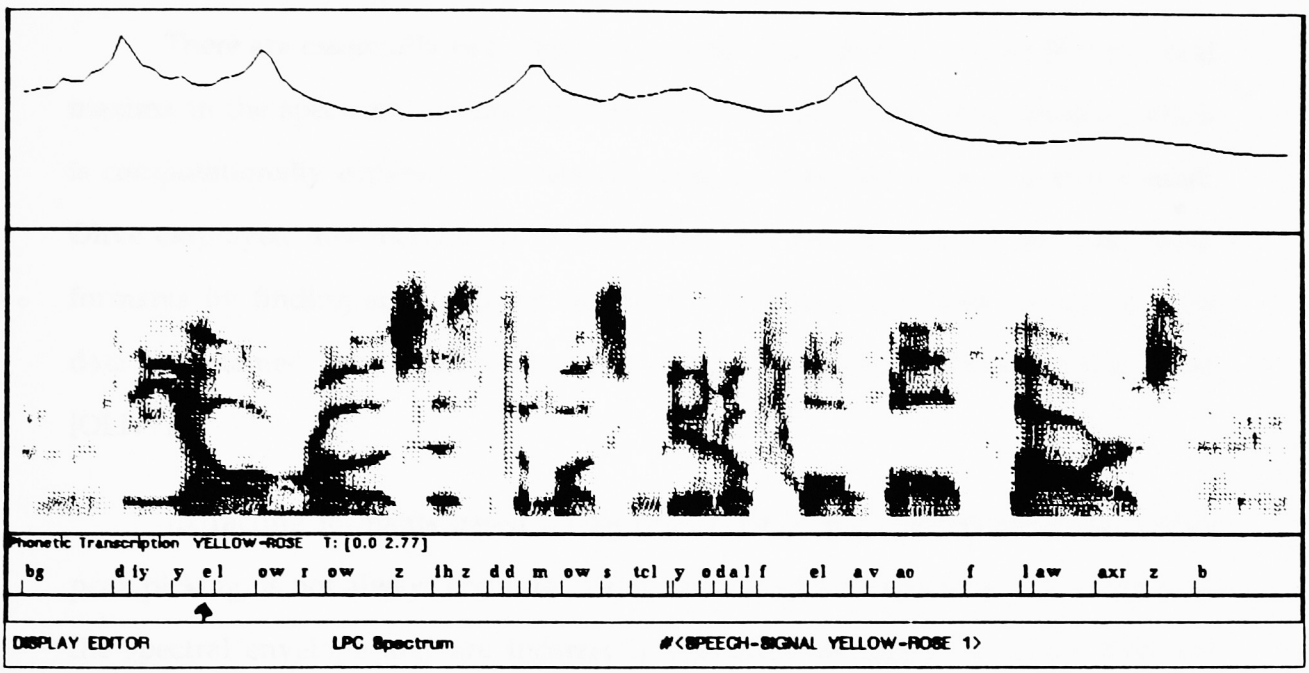


Figure 2-18
LPC Spectrum for a Frame in phoneme /eh/ and the Corresponding Spectrogram

Linear Predictive Coding (LPC) was first applied to speech in the early 1970s. It is one of the most heavily used techniques for producing the spectra from which formant frequencies are derived. LPC is based on a source filter theory model of the vocal tract. The smooth spectrum that is derived through LPC analysis corresponds

to the vocal tract frequency response curve, and peaks in the LPC spectrum are closely associated with vocal tract resonances or formants (see Figure 2-18).

In speech, linear prediction can be related to the cylindrical model of the vocal tract. There are three possible models: mixed pole-zero, all-pole, and all-zero. The all-pole model is most frequently used in speech because it is the easiest to compute and, if one ignores the nasals and some of the fricatives, the transfer function is an all-pole function [PARS86].

There are essentially two ways to proceed: compute the roots or find the local maxima in the spectrum envelope derived from the predictor. Root-finding, which is computationally expensive, involves finding the complex roots of a polynomial. Olive employed this method to extract formants He determined the first three formants by finding simultaneous solutions to the least-square-fit equations. The data he obtained were used to control a computer-simulated formant synthesizer [OLIV71].

Extracting formants based on local maxima in the spectral envelope, called peak picking, is not always accurate. Often various smaller spurious peaks appear in the spectral envelope and are incorrectly identified as formants. Also formants sometimes come so close together that they may appear to be one formant. McCandless [MCCA74] applied an algorithm to determine which of the peaks correspond to the first four formants. The algorithm determines an anchor point in the middle of a high energy voicing (vowel) area. From this anchor the algorithm works forward in time until a low energy voicing area is encountered. Then the algorithm works from the anchor backward in time until a low energy voicing area is again reached. For each slice of time the following six steps are performed.

1) The first four peaks with frequencies between 150 and 3400Hz become the peak values.

2) The four formant slots are filled with a peak based on the absolute smallest difference between the frequency of the peak and frequency of the estimate for that slot.

3) If a peak is used in more than one slot, it is removed from the slot whose absolute difference between the frequency of the peak and the frequency of the slot's estimate is larger.

4) If there is an unassigned peak, assign it to its corresponding slot if it is empty. If the slot is not empty, assign the peak to the slot if the amplitude of the peak is twice as large or larger than the amplitude of the peak in currently in the slot position. If the peak is still not assigned and slot + 1 is empty, move the peak currently in the slot to slot + 1 and assign the peak value to slot. If the peak is still not assigned and slot - 1 is empty, move the current slot value to slot - 1 and assign the peak to the slot.

5) If there are any unfilled slots, recompute the spectrum on a circle with a radius less than one to enhance the formants, and thereby separating two merged formants. Start over with step 1.

6) The answers (filled slots) are recorded as formants for this frame and as estimates for the next frame.

Markel extracted spectral peaks using an algorithm based on the inverse filter which attempts to transform the input signal into a constant or white noise spectrum. As it approaches infinity, the inverse filter theoretically predicts the exact inverse of the input spectrum, resulting in a constant for the output or error spectrum. He then assigned formant labels to the peaks according to a set of rules [MARK72].

Another method of tracking formants is a statistical approach using hidden Markov models and vector quantization [KOPE86]. Kopec used a large training set to determine the likelihoods that a formant would be in particular frequency ranges for particular vector quantities. These, in turn, were used as the observation probabilities in the hidden Markov models. A hidden Markov model is

characterized by a 5-tuple. Each tuple is described as it pertains to Kopec's formant tracker.

1. a set of states which corresponds to the possible formant values;
2. a set of symbols which consists of an LPC vector quantization codebook and time, the analysis frame index in a short-time LPC vector quantization of the speech waveform;
3. a vector of prior probabilities;
4. a matrix of transition probabilities in the formant tracker model may be viewed as expressing continuity constraints on formant motion; and
5. a matrix of observation probabilities which expresses the relationship between the observed short-time speech spectrum and the formants of the underlying vocal tract configuration.

Detection and estimation are performed simultaneously by including a constant which represents the absence of a formant. Formants can be tracked singly by using scalars for each state or jointly by using vectors for each state [KOPE86].

Gayvert also used a statistical approach. He tracked formant trajectories in continuous speech by applying a probability measure to a set of features extracted from each analysis frame of the speech signal and using a conditional mean estimate to determine formant frequency values. The features used can be vector quantization symbols, spectrum levels, or other sets of features related to formant frequencies. Continuity constraints can be applied by using simple smoothing algorithms or hidden Markov models [GAYV89].

CHAPTER 3

PROJECT IMPLEMENTATION

The goal of the system is to examine the peaks in the Linear Predictive Coding (LPC) files to determine which peaks correspond to F1, F2, and F3. This determination is based on the sex of the speaker, the target phoneme, the phonemes on either side of the target phoneme, and the starting and ending times for the target phoneme.

3.1 Data Structures

When a phoneme is processed, any number of phonemes on either side of that phoneme may be examined. This group of phonemes, referred to as a phoneme window, is an array of structures in which each structure represents one phoneme. In the present implementation, only three phonemes are examined at any one time: the previous phoneme, the target phoneme, and the next phoneme. After the target phoneme is processed, the formant frequencies for the previous phoneme are written to an external file. The phoneme window is then moved to the right (see Figure 3-1).

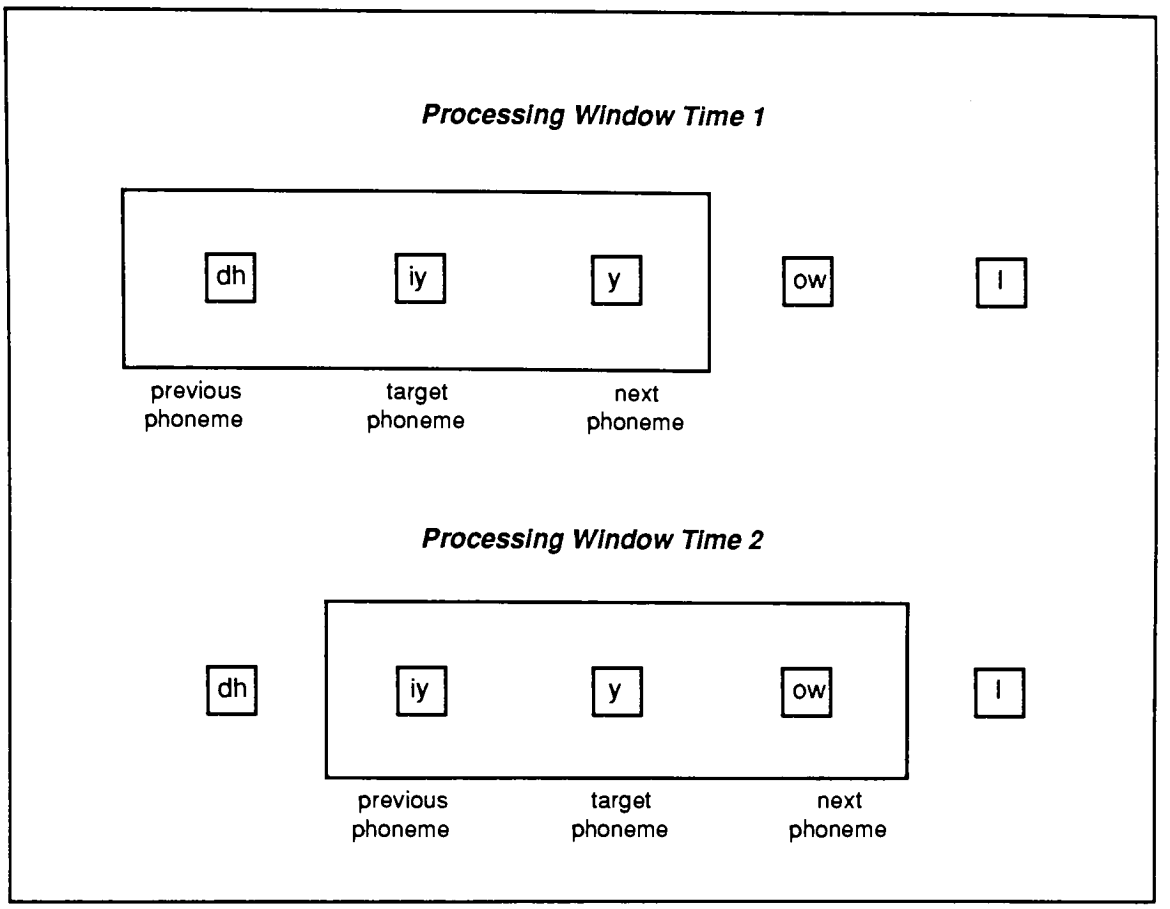


Figure 3-1
Movement of the Phoneme Window in the Domain of Time.

Each structure contains the original LPC peaks and amplitudes, the peaks the system has decided are the first three formants, the peaks it has not used, the estimated values, the maximum expected formant values, and the minimum expected formant values for that phoneme. (The procedures used to derive estimated formant frequency values and their corresponding maximum and minimum values will be described in section 3.4.1.) It also contains the phoneme group to which the phoneme belongs and the phoneme's starting and ending frames.

3.2 Phoneme Groups

Each phoneme is classified into one of the groups listed in Table 3-1.

Group	Example
nonvocalic	g
monophthong	eh
diphthong	ey
semivowel	l
nasal	n

Table 3 - 1
Phoneme Groups Used in the Current Study

Monophthongs, diphthongs, semivowels, and nasals have a resonant quality because they are typically voiced and produced with a relatively open vocal tract. That also means they have well-defined formant patterns and are, therefore, of interest to this study. All other phonemes are labeled nonvocalic and essentially ignored. If the target phoneme is a nonvocalic phoneme, the phoneme window is immediately shifted to the right. If a nonvocalic phoneme is in the position of the previous or next phoneme, the coarticulatory effects on the target phoneme are not taken into account for the current study

Presently nasals are not labeled by the system. It is difficult to determine if the peaks in a nasal phoneme are actual formants because of antiresonance. They do

influence adjacent phonemes and, therefore, expected formant frequency values are maintained for the first three formants for all nasals.[3]

Diphthongs represent a unique case because they either have two steady states connected by a transition or are entirely transitional. For this reason, they are processed as if they were two separate phonemes and are later joined. Semivowels are transitions with a brief or nonexistent steady state portion. At the start of this project, it was assumed semivowels would be processed uniquely because of the rapid movement. However, this proved to be unnecessary, and they are currently processed in the same way as monophthongs. However, when continuity between phonemes is checked, some combinations of semivowels and vowels appear to the computer to be discontinuous because of this rapid change of formant frequency values. Therefore, there are some combinations of semivowels and vowels that are not considered when checking for continuity between phonemes.

---

[3]See Appendix B for estimated expected formant frequency values.

3.3 Test Data

The test data for the system came from C-MU. For this study eight speakers were selected, four males and four females. Each speaker produced ten sentences comprising 1,011 resonant phonemes which cover 17,363 5 msec. frames. The LPC file contains the frequencies and amplitudes of the peaks in the LPC spectrum. Figure 3-2 describes the signal processing steps taken to obtain the current input to the system from the raw speech signal.

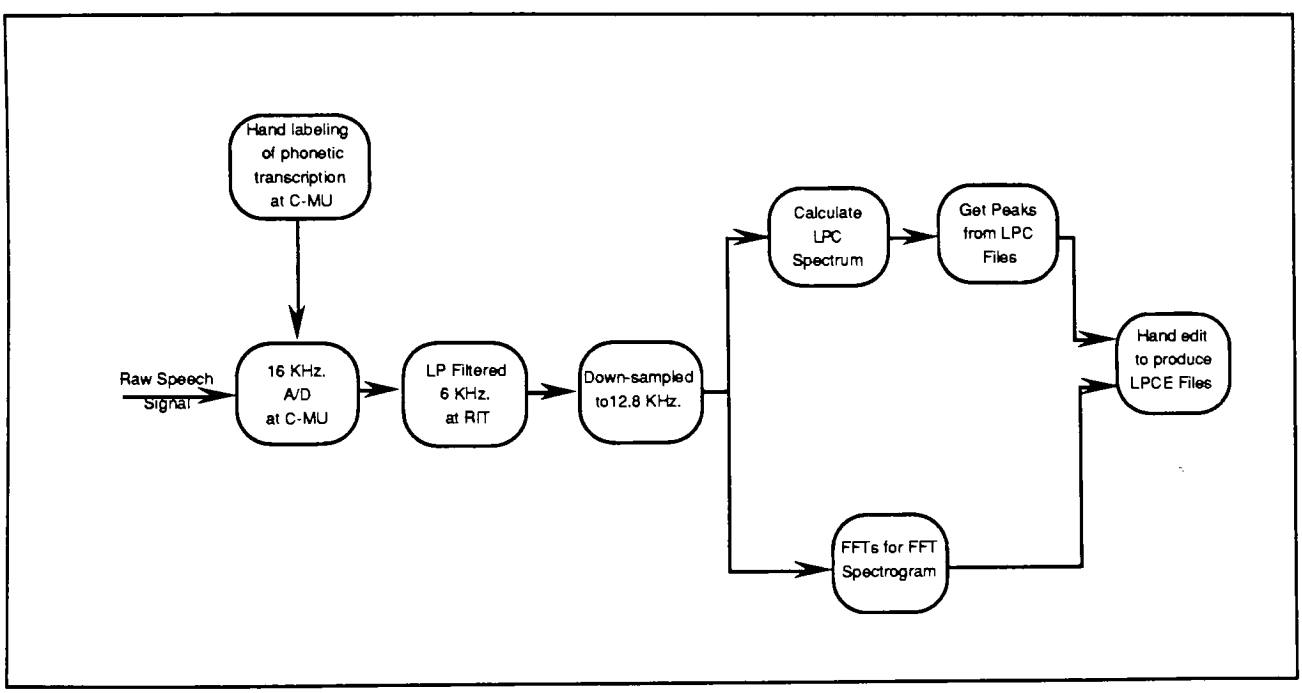


Figure 3 - 2
Diagram of How the Raw Data is Processed Prior to Use by this System

The LPC is calculated every 5 msec. Figure 3 - 3 is an ASCII representation of a portion of an LPC file. The first column is the frame number; the second column

is the first peak (P1); the third column is P1s corresponding amplitude (A1); the remaining columns are the peaks two through seven each followed by its corresponding amplitude. The corresponding Label file contains the hand-labeled phonetic transcription and the starting times for each phoneme.

FRAME #	P1	A1	P2	A2	P3	A3	F4	A4	P5	A5	P6	A6	P7	A7
1640	899	9585	1891	9312	3025	9018	3324	5807	3967	9179	4971	9550	0	0
1645	884	9604	1722	9193	2926	6337	3031	9292	4078	8956	5002	9609	0	0
1650	840	9669	1715	9456	2606	6581	3050	9262	3951	8913	5013	9583	0	0
1655	775	9705	1325	8003	1866	9482	2952	9286	3783	9154	4913	9293	0	0
1660	760	9841	1441	3262	1924	9489	2812	9120	3834	8856	4854	9333	0	0
1665	752	9878	1704	8504	1961	8833	2866	9175	3745	8808	4891	9230	0	0
1670	742	9879	1883	3887	1971	8150	2871	9274	3717	9016	4897	9205	0	0
1675	727	9898	1921	9071	2085	7928	2864	9321	3730	9178	4838	9043	0	0
1680	708	9923	1937	3789	2177	8426	2886	9301	3705	9290	4763	9089	0	0
1685	692	9943	2110	9010	2163	8026	2925	9200	3780	9259	4756	9079	6348	7086
1690	684	9943	2164	9299	2270	7033	2910	9150	3777	9240	4735	9247	0	0
1695	682	9896	2184	9238	2455	5946	2854	9239	3752	9252	4753	9304	0	0

Figure 3 - 3
Ascii Representation of a Segment of an LPC File
where P1 through P7 are the First Seven Peaks and A1 through A7 are the Peaks' Corresponding Amplitudes. Frame Number Refers to the Time Segment.

After displaying the spectrogram, the plot of the LPC peaks, and the labels, the expert would trace over the LPC peaks using LPC Tool. LPC Tool is a graphics tool for the SUN written by Eric Luce to edit a plot of the LPC peaks. Using the mouse, peaks can be deleted, interpolated, or traced. The output is an LPC peak file. The hand-edited LPC files are used to test the performance of the system.

3.4 System Architecture

In order to determine which of the LPC peaks are formants, three distinct steps are followed: 1) A starting frame is selected; 2) Continuity between the frames in a phoneme is checked; and 3) Continuity between adjacent phonemes is checked.

3.4.1 Selection of a Starting Frame

Selecting a good starting frame is crucial to the success of the system. From the starting frame the system determines which peaks in the adjacent frames are formants, and these choices influence decisions in the following frame. The algorithm continues in this manner until the last frame is reached. This process is followed from the starting frame back in time to the first frame. If peaks are chosen that are actual formants in the starting frame, it is much more likely to continue labeling the formant correctly.

The starting frame is selected on the basis of two criteria: 1) the location of the frame within the phoneme, and 2) the frequency of the peak in relation to expected values for the phoneme.

A frame in the center of the phoneme is a better choice than frames closer to either boundary because the effects of coarticulation are reduced. LPC peaks that fall within accepted ranges for the target phoneme are more apt to be the actual formant values and therefore a good starting place.

To pick a starting frame intelligently, the system must know the identity of the target phoneme, the identity of adjacent phonemes, and the sex of the speaker. The system has knowledge of the expected ranges for individual phonemes and how

particular adjacent phonemes influence that phoneme. It is also necessary to know the starting and ending frames to determine the central area of the phoneme.

The estimated values for each formant are based on the sex of the speaker and the target phoneme. Most of these values are from Peterson and Barney [PETE52] and Allen et al. [ALLE87] (see Appendix B). Three phonemes, /oe/, /o/, and /e/, were not included in either study. The estimated values for /oe/ were derived from a portion of the C-MU data The phoneme /e/ is a nondiphthonized version of /ey/. The estimates for /eh/ were used as a basis, adjusting F1 down and F2 upward. The phoneme /o/, a nondiphthonized version of /ow/, is similar to /uw/. Using the estimates for /uw/ as a basis, F1 and F2 were adjusted upward.[4] Neither /e/ nor /o/ occur frequently . In fact, there were no occurrences of /e/ in the test data for this project.

The phonemes used in the Peterson and Barney study [PETE52] have data for both male and female speakers, as well as corresponding maximum and minimum values. The values for the remaining phonemes, those taken from Allen et al. [ALLE87] and from the C-MU data, had no data for females and/or no corresponding maximum or minimum values associated with the estimates for either males or females. To derive a value for female speakers, each estimate not covered in the Peterson and Barney study was increased 17% over its male counterpart.[5] A formula was derived from the maximum values and the estimates of the Peterson and Barney data [PETE52] using linear regression. Another formula was similarly

---

[4]This information was obtained from James Hillenbrand.

[5]The value 17% was suggested by Fant [FANT73].

created using the minimum values and the estimates from the Peterson and Barney data [PETE52]. These two formulas, one for the maximum values and the other for the minimum values, were used to estimate maximum and minimum values for phonemes not included in the Peterson and Barney study.

3.4.2 Continuity within the Phoneme

After determining a starting frame, a base point for each of the three formants, the system implements an algorithm to decide which points are most likely to correspond to the first three formants. This algorithm is based on the one developed by McCandless [MCCA74] for tracking formants. The current system differs from McCandless in that the phonetic transcription, the starting time, and the ending time for each phoneme is known. The purpose of the current system is to incorporate knowledge about expected values for particular phonemes in deciding which LPC peaks correspond to formant frequencies.

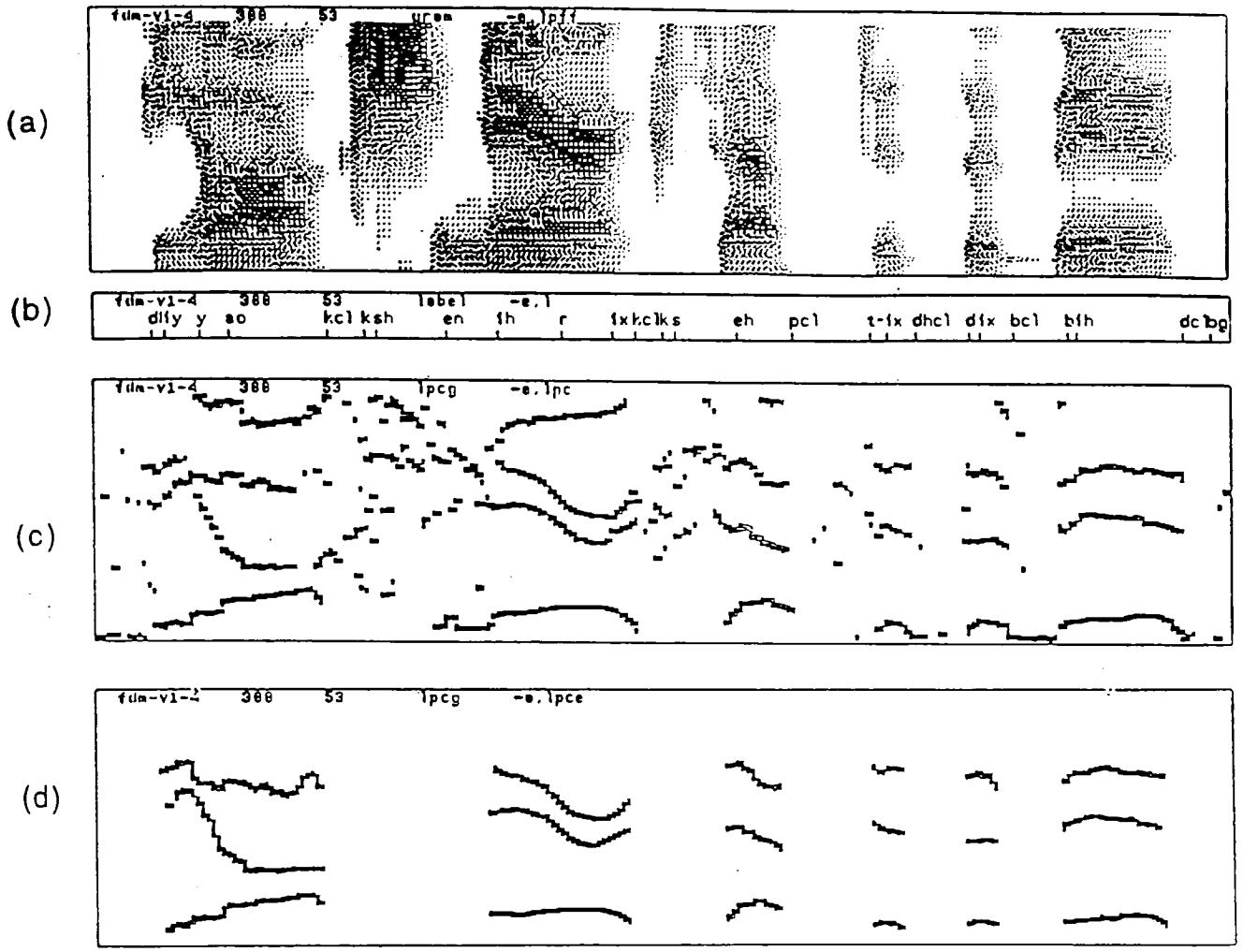


Figure 3 - 4
Example of a Spectrogram, a Plot of LPC Peaks, and the Label File an Expert Would Use to Trace the Formants.[GAYV89]

For each frame the four formants of the phoneme are processed at one time using the arrays "slot", "peak", and "estimate". Slot is a two-dimensional array which contains the peaks the system believes are formants. One dimension is time; the other dimension is formant frequency (F1, F2, and F3). Peak is also a two-dimensional array. One dimension is time; the other is the peaks 1 through 7, referred to as P1, P2, P3, P4, P5, P6, and P7. The slots (initially an empty array) are filled with peaks (the frequency values of LPC peaks from the input file). The decision as to which peaks should occupy each slot is based on an estimate for each

of the formants (referred to in the diagrams as est1, est2, and est3 for the estimates of F1, F2, and F3 respectively. The first estimate values are derived from information in the database (referred to as the original estimate) and the system's knowledge about the coarticulatory effects of the adjacent phonemes.

Algorithm

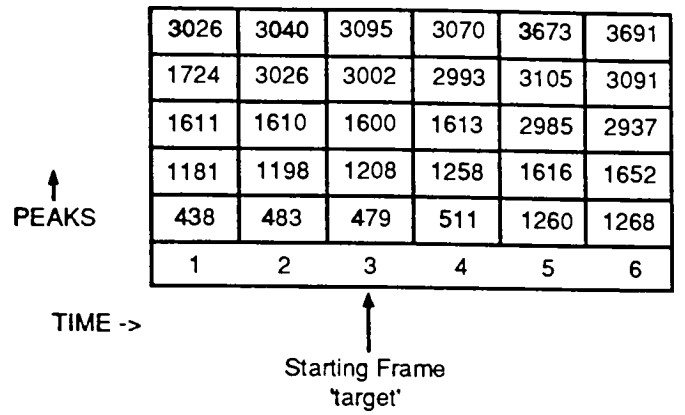
```
fill_slots_for_the_first_time
        fill each slot with a peak closest to its corresponding estimate

remove_duplicates
        the slot whose estimate is closest to that peak retains that peak; the other slot is set
        back to zero

 fill_slots_for_the_second_time
        if a  slot is empty and the corresponding peak is not, then
                if the 'target' peak > 'previous' slot & the 'target' peak < next slot then
                fill slot with peak
        if it is the 'start' frame then
                if slot is empty and peak + 1 is unused then
                        if peak + 1 > previous slot and peak + 1 < next slot then
                                fill slot with peak + 1
                if slot is empty and peak - 1 is unused then
                        if peak - 1 > previous slot and peak - 1 < next slot then
                                fill slot with peak - 1
```

NOTE: If the difference between the peak and the estimate is greater than a threshold defined as MAXDIFF, the slot will not be filled by the peak. The slot will remain 0, and the estimate for the 'next' frame will remain the same value it is for the current frame.

Figure 3-5
Algorithm to Fill Slots with Available Peaks

The algorithm, described in Figure 3-5, consists of three steps: filling the slots for the first time, removing duplicates, and filling the slots for the second time. Figure 3-6 (a) shows the values for the estimates, peaks, and slots at the beginning of 'start' frame. In Figure 3-6 (b) the slots are filled with peaks for the first time. The selection is based on the smallest absolute difference of the estimate for that formant and each of the first four peaks. In Figure 3-6 (c) duplicate values are removed. The

slot whose corresponding estimate is the closest (smallest absolute difference) retains the peak; the other slot(s) are set to zero. In Figure 3-6 (d) the slots are filled for the second time. If there are no empty slots, the frame is considered done. If a slot is empty and its corresponding peak has a value, the peak is assigned to the slot under the following conditions: 1). the peak value is greater than any lower slot value; 2). the peak value is smaller than any higher slot value, or 3). the absolute difference between the estimate for the slot and the peak value is less than MAXDIFF, a programmer defined threshold. If the slot is still empty and it is the first frame, the other peaks are checked. Before a peak is transferred to a slot, it must meet the above three conditions.



(a) Starting Frame values at Beginning of 'Start' Frame

Estimates from database	est1: 330	est2: 1060	est3: 1380		
Peak values	P1: 479	P2: 1208	P3: 1600	P4: 3002	P5: 3095
Slot Values	S1: 0	S2: 0	S3: 0		

(b) Starting Frame values after Filling Slots for the First Time

Estimates from database	est1: 330	est2: 1060	est3: 1380		
Peak values	P1: 0	P2: 0	P3: 1600	P4: 3002	P5: 3095
Slot Values	S1: 479	S2: 1208	S3: 1208		

(c) Starting Frame values at after Remove Duplicates

Estimates from database	est1: 330	est2: 1060	est3: 1380		
Peak values	P1: 0	P2: 0	P3: 1600	P4: 3002	P5: 3095
Slot Values	S1: 479	S2: 1208	S3: 0		

(d) Starting Frame values at after Filling Slots for the Second Time

Estimates from database	est1: 330	est2: 1060	est3: 1380		
Peak values	P1: 0	P2: 0	P3: 0	P4: 3002	P5: 3095
Slot Values	S1: 479	S2: 1208	S3: 1600		

Figure 3 - 6
Description of Filling Slots with Peaks

The slot values become the formant values for the current frame; they also become the new estimates for the next frame as the selection of peaks continues. This process continues until the last frame of the formant is reached. Then the estimate becomes the original estimate and the process continues from the start frame minus one

frame to the first frame of the formant. Continuity of formant frequency values is checked at the boundary of phonemes. (For a description of the process see section 3.4.3.) The phoneme window then slides one phoneme to the right.

The first or 'starting' frame of each phoneme is processed somewhat differently from the rest of the frames in the phoneme. At the outset there are no values in the slot array to compare against; the estimate for the starting frame is from the database. This estimate is an average over all speakers for that sex. For all other frames, the estimate is a frequency value from the LPC file. Therefore, there is no threshold for a first frame; a 300 Hz threshold is enforced for all other frames. The starting frame is chosen by comparing the peak values in each frame to the formant frequency maximum and minimum values in the database and by examining the location of the frame within the phoneme. Each frame is assigned points for each formant that is within the boundaries of the maximum and minimum values and for being located in the center half of the phoneme. The starting frame that is selected is the one that has the most points. If there is a tie(s), a frame closer to the center is selected.

```
1       For each phoneme
2               {
3               determine the phoneme group [6]
4               get the (original)[7] estimate for that phoneme
5               get the max and min limits
6               determine a good "start" frame for  the phoneme
7               from "start" to  the last frame
8                       {
9                       fill slots with best choice from peaks [8]
10                      estimates for next frame are the current  slot values
11                      }
12              estimate becomes original  estimate
13              for each frame from "start" - 1 to the first frame
14                      {
15                      fill slots with best choice from peaks
16                      estimates for next frame are the current slot values
17                      }
18              }
```

Figure 3-6
Pseudocode Describing the Method of Determining which Peaks are Formants

---

[6]A phoneme group would be vowel, diphthong, semivowel, nasal, or nonvocalic.

[7]The original estimate is used for the 'start' frame and the 'start' frame - 1. It refers to the estimate retrieved from the database.

[8]This process is described in Figure 3 - 5.

3.4.3 Continuity between Phonemes

If two adjacent phonemes both have formants, then the formant frequency values should not show abrupt discontinuities at the boundary between phonemes. In this study two adjacent formants are considered continuous

1. if the average of the five boundary frames (frames containing zero values are excluded) for one formant is within 300 Hz of the average of the five boundary frames for the other formant,

2. if the absolute difference between each of the three boundary frames is less than 300 Hz, or

3. if either one or both of the formants contain zeros at each of the five boundary frames.

For each formant that is not continuous (F1, F2, and F3), the unused peaks are checked for suitable peaks. A peak is considered suitable if it is within 300 Hz of the average of the five boundary frames of the other formant. If there are suitable unused peaks for one phoneme (for this example 'target' has suitable unused peaks). the slot values for 'target' for that particular formant are moved to the unused peaks, and that formant is reprocessed using the values from 'previous' as an estimate. If both have suitable unused peaks, the phoneme that has the lowest percentage of suitable unused peaks is used as the base phoneme, and the other other phoneme is reprocessed. If neither phoneme has suitable unused peaks, the other formants are checked for continuity. If there are other formants that do not match, these are processed. The original formant is rechecked because other peak values may have been placed in the unused category. If there are still no suitable unused peaks, the original values are retained.

3.4.4 Adjustments for Diphthongs

In this study a diphthong is considered as two separate vowels, each part different yet from its monophthong counterpart, to be tracked separately and then joined. The first part of the phoneme is processed using the front two-thirds of data; the second part is processed using the back two-thirds of data. The center third of data for both should be identical. If it is, the selected peaks are probably the formants. If the center third does not match, the last frame in the front two-thirds of data is used as the estimates for F1, F2, and F3. The last third of the data is processed processed in the usual way.

3.4.5 Adjustments for Semivowels

Semivowels are shorter in duration than vowels, and therefore usually do not have a steady state. This makes them more difficult to track. To determine continuity within the phoneme, the semivowel phoneme is processed in the same way as vowels. To determine continuity between phonemes, the semivowel, because of its short duration and transition, is not always recognized as being continuous with its neighbor, especially if the expected formant frequency values of the two phonemes are distant from each other. Therefore, the following combinations of semivowels and vowels or vowels and semivowels are are not checked for continuity at the phoneme boundary:

/w/	/iy/
/iy/	/w/
/iy/	/y/
/iy/	/r/
/r/	/ih/
/r/	/ix/
/ix/	/r/

CHAPTER 4

RESULTS

The results reported in this section are based on four male and four female speakers, each of whom read ten sentences. The first version follows the steps outlined in sections 3.4.1 - 3.4.3 This version is referred to as the "matched" version because boundary values of adjacent phonemes are matched or compared to see if they are continuous. The second version, which is the same as the first version except that it doesn't implement the matching routine described in section 3.4.3, is referred to as the "unmatched" version. Unless there is a specific reference to a version, the matched version is being discussed.

4.1 Performance Evaluation

The output of the system is measured against the hand-edited files, which are assumed to be correct. Research has shown that hand-editing of formant frequencies is accurate within about 50 Hz [MONS83]. Chapter 5 discusses some problems involved with this assumption. The statistics gathered are false alarms, misses, mean absolute error, percent large errors, RMS error, percent matched, and error histograms.

A formant in a frame is considered a miss if the system has a value of zero and the hand-edited file has a formant value. A false alarm, on the other hand, is when the system has a formant value and the hand-edited file has a value of zero. A false alarm is usually not an error on the part of the system. It means that there is a peak value in the LPC file within the specified limits. False alarms usually occur in

one of two places: at the boundary of a phoneme (usually at the end) or at the end of an utterance. At the boundary there may be a discrepancy where the formant starts and stops. The system is working with starting and ending values for each phoneme, which are provided in the label file. The expert is working with the spectrogram, the label file, and the plot of the LPC peaks using LPC Tool on a monitor. The resolution is not as exact as discrete points and, consequently, discrepancies may occur as in Figure 4 - 1. At the end of an utterance a speaker's voice may trail off making the formants weak. It is not uncommon to see entire phonemes at the end of an utterance in the hand-edited file all zeros (see Figure 4 - 2). Misses are not always errors, but are more likely to be errors than false alarms. In Figure 4 - 3 there are 3 misses at the end of F3. The difference between F3 peak 2179 in the system file and the best peak choice for the next frame in the LPC file 2663 is greater than the threshold (a programmer defined constant) set.

Values in LPC File					Hand-edited File			System File		
498	1200	2268	3216	3514	498	1199	2266	498	1200	2268
502	1218	2316	3210	3513	501	1214	2310	502	1218	2316
498	1245	2339	3226	3504	498	124	2342	498	1245	2339
481	1305	2278	3139	3468	481	1302	2279	481	1305	2278
454	1312	2274	3168	3442	454	1311	2219	454	1312	2274
424	1310	2219	3065	3409	426	1307	2219	424	1310	2219
329	1321	2193	3312	3626	0<-	0<-	0<-	329	1321	2193
289	1333	2112	3207	3787	0<-	0<-	0<-	289	1333	2112
288	1288	2177	3124	3851	0<-	0<-	0<-	288	1288	2177

Figure 4 - 1
Example of False Alarms in phoneme /ax/ at Boundary

Values in LPC File					Hand-edited File			System File		
545	1512	2223	2764	716	0<-	0<-	0<-	545	1512	2223
395	1478	2212	2778	3844	0<-	0<-	0<-	395	1478	2212
598	1486	2183	2851	3421	0<-	0<-	0<-	598	1486	2183
606	1489	2198	2873	3499	0<-	0<-	0<-	606	1489	2198
585	1488	2209	2889	3621	0<-	0<-	0<-	585	1488	2209

Figure 4 - 2
Example of False Alarms in phoneme /ax/ at the End of an Utterance

Mean absolute error is based on the sum of the absolute differences between the target value in the system and the corresponding target in the hand-edited files. This sum is then divided by the total number of frames. RMS error, the root mean square error, is the standard deviation of the signed difference between hand-edited and system formant frequency values. Theoretically the mean absolute error and RMS error should be zero if a phoneme is labeled correctly, because input to both systems came originally from the same data. As can be seen in Figure 4 - 4, there are no LPC peaks that correspond closely to the hand-edited choice of peaks for F3. Some differences are on account of the smoothing of the hand-edited files and the interpolation performed by LPC Tool. For this reason, mean absolute error and RMS error are somewhat difficult to interpret; these error measurements will be influenced partly by system errors, but also by inherent and unavoidable discrepancies between formant frequency values in the hand-edited files and peak values in the raw LPC files.

Values in LPC File					Hand-edited File			System File		
487	1307	1860	3175	0	486	1298	1817	487	1307	1860
483	1310	1659	3160	3890	485	1304	1837	483	1310	1659
493	1340	1541	3161	3511	495	1342	1823	493	1340	1541
486	1320	1602	3173	3537	487	1327	1803	486	1320	1602
491	1355	1557	3180	3327	494	1368	1783	491	1355	1557
482	1328	1607	3191	3316	485	1347	1763	482	1328	1607
487	1323	1608	3186	3327	491	1338	1743	487	1323	1608
478	1330	1621	3193	3356	481	1350	1723	478	1330	1621
478	1353	1621	3186	3253	486	1380	1703	478	1353	1621
476	1364	1647	3187	3295	481	1390	1683	476	1364	1647
472	1384	1657	3186	3212	473	1400	1663	472	1384	1657
466	1399	1661	3155	3200	469	1410	1668	466	1399	1661
455	1416	1664	3010	3220	468	1450	1686	455	1416	1664
450	1448	1702	3195	3213	463	1468	1766	450	1448	1702
443	1466	1713	3176	3189	458	1509	1846	443	1466	1713
437	1512	1765	3168	3380	441	1509	1926	437	1512	1765
433	1554	1877	3157	3420	436	1609	2006	433	1554	1877
432	1602	2179	3149	3166	435	1709	2265	432	1602	2179
426	1625	2663	2890	3160	427	1843	2610	426	1625	0 <-
418	1702	2663	3161	0	420	2038	2616	418	1702	0 <-
414	1738	1855	2776	3177	416	2213	2700	414	1738	1855
404	1519	2637	3208	0	407	2272	2723	404	1519	0

Figure 4 - 3
Example of Misses in phoneme /r/

Percent large error is a measurement defined by Kopec [KOPE86] and used by Gayvert [GAYV88]. Kopec considers errors larger than 250 Hz for F1 and larger than 500 Hz for F2 and F3 large errors. It is used in this study as a means of comparison to Kopec and Gayvert.

For each frame the routine to determine percent correct examines each formant value (F1, F2, and F3) in the hand-edited files, and selects the closest value in the corresponding frame of the LPC file. If the chosen value from the LPC file is identical to the system's selected formant value (the formant corresponding to the formant under consideration in the hand-edited files), then the system selected

formant for that frame is considered to be correct. The last measurement made is an error histogram. The absolute difference between the hand-edited value and the system value for each formant is calculated. Each division is 100 Hz, except the last division which includes all errors over 900 Hz.

Values in LPC File					Hand-edited File			System File		
368	1115	2216	3637	0	375	1156	2685	368	1115	2216<-
340	1119	2170	3647	0	345	1138	2690	340	1119	2170<-
333	1093	2289	3636	0	331	1091	2695	333	1093	2289<-
338	1107	2389	3621	0	338	1101	2700	338	1107	2389<-
335	1151	2222	3620	0	336	1155	2705	335	1151	2222<-
353	1201	2171	3600	0	354	1219	2710	353	1201	2171<-
383	1219	2148	3571	0	382	1225	2715	383	1219	2148<-
391	1210	2246	3494	0	387	1251	2720	391	1210	2246<-
420	1297	2600	2857	0	421	1331	2724	420	1297	2600
429	1437	2776	3889	0	432	1446	2768	429	1437	2776
421	1480	2737	3856	0	424	1485	2729	421	1480	2737

Figure 4 - 4
No LPC Peak in /l/ that Corresponds to Hand-edited Choice for F3

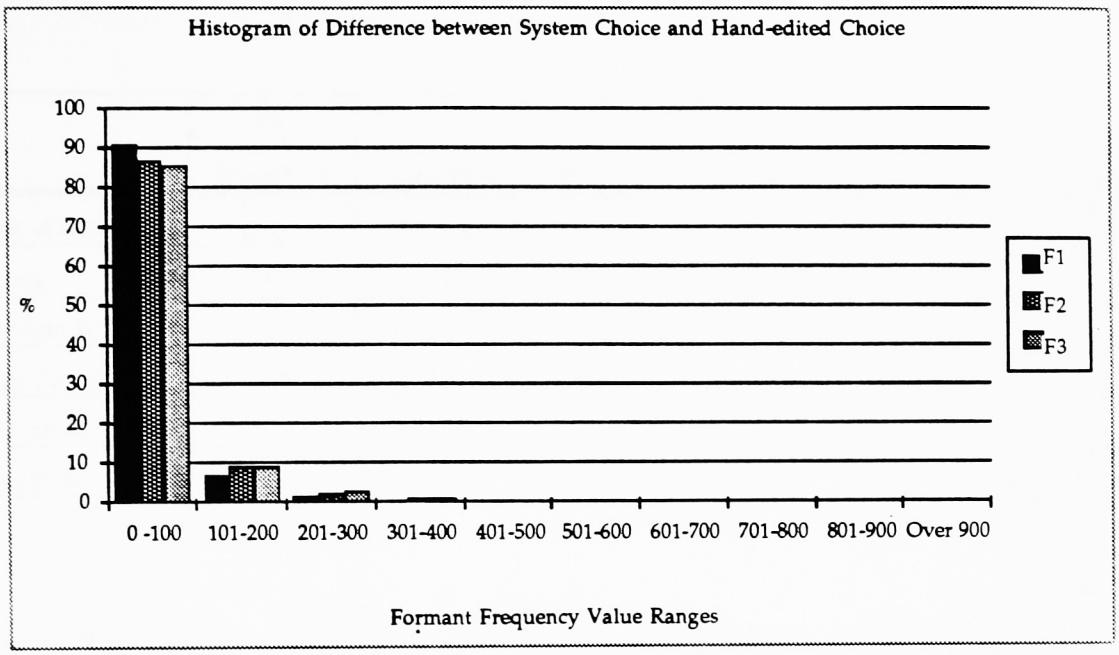


Figure 4 - 5
Error Histogram for All Phonemes in Study
Note: Errors below 1% did not show up on Bar Graph

4.2 Performance

	F1 Matched	F1 Unmatched	F2 Matched	F2 Unmatched	F3 Matched	F3 Unmatched
False Alarms:	11.3%	11.3%	12.9%	13.0%	17.1%	17.0%
Misses:	1.0%	.9%	2.6%	2.3%	4.2%	4.7%
Absolute Error:	32	32	69	75	94	104
Large Error:	.4%	.5%	3.4%	3.9%	5.8%	6.8%
RMS Error:	90.9	88.73	210.29	222.19	294.92	319.48
Percent Matched:	98.9%	98.6%	92.2%	91.3 %	88.8%	87.4%

Table 4-1
Results Tabulated for All Speakers Using the Matched Version and the Unmatched Version. Data Consisted of a Total of 1, 011 Phonemes Covering 17,363 Frames.

The results averaged across all speakers are presented in Table 4-1. Looking at the matched version, the system correctly selected 98.8% of the peaks for F1, 92.2% of the peaks for F2, and 88.8% of the peaks for F3. The general trend for both the matched and unmatched versions is the lower the formant, the better the results. The matching segment examines the selected peaks in the 'previous' phoneme and the 'target' phoneme as well as the unused peaks for both these phonemes in the LPC file. It sometimes selects the wrong peak values to retain as formants, thereby changing peaks that were correct. However, the slightly better performance for the matched version indicates that the match routine corrected more frames than it mislabeled. For some phoneme groups, such as glides, the match routine significantly improved the results (see Table 4-2).

	F1	F1	F2	F2	F3	F3
	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched
Percent Matched	96.6	96.8	90.5	86.9	86.9	81.3

Table 4-2
Comparison of Results for Glides Using the Matched Version and the Unmatched Version

	F1 Male	F1 Female	F2 Male	F2 Female	F3 Male	F3 Female
False Alarms:	13.4%	9.6%	15.9%	10.5%	22.6%	12.5%
Misses:	0.7%	1.3%	16.7%	3.1%	16.7%	5.7%
Absolute Error:	22	39	45	81	110	81
Large Error:	1.0%	1.9%	1.9%	3.9%	8.1%	4.0%
RMS Error:	83.9	95.2	172.8	217.8	351.3	240.9
Percent Correct:	99.0%	98.8%	95.9%	90.6%	85.3%	91.4%

Table 4-3
Comparison of Results for Male and Female Speakers

The system performed significantly better (95.1% versus 89.8%) on the males than the females for F2. The reverse is true for F3: the system identified 91.5% of the F3 formants for the female speakers versus 85.2% for the male speakers (see Table 4-3).

Results are also tabulated according to the following phoneme groups:

Front Vowels	iy, ih, eh, ae, e
Central Vowels	ix, ax, ah
Back Vowels	ux, uw, ao, o, uh, aa, oe
Diphthongs	ey, ay, oy, aw, ow
Retroflex Vowels	er, axr
Liquids	r, l
Glides	y, w
Syllabic Resonants	el, em, en

	F1	F2	F3
FRONT VOWELS	99.0%	88.8%	89.5%
CENTRAL VOWELS	97.8%	96.2%	92.6%
BACK VOWELS	99.6%	98.7%	93.9%
DIPHTHONGS	99.6%	91.2%	84.6%
RETROFLEX VOWELS	98.4%	91.2%	90.9%
LIQUIDS	99.2%	95.3%	89.7%
GLIDES	96.8%	91.1%	87.1%
SYLLABIC RESONANTS	98.3%	97.7%	75.9%

Table 4-4
Percentage Correct for All Speakers Tabulated according to Phoneme Groups

Looking at the Front, Central, and Back Vowels for F2, in Table 4-4 the performance improves with movement from the front to the back. Worth noting is that the expected formant frequency values for F2 decrease with movement from the back to the front. The system performed poorly on F3 for syllabic resonants. The F3 peaks selected by the hand-edited version usually differ greatly from the estimated expected formant frequency values (see Figure 4 - 7).

Values in LPC File					Hand-edited File			System File		
177	1204	2196	2999	3864	375	1379	1390<-	177	1204	2999
345	1401	2274	2787	3693	375	1379	1390<-	345	1401	2787
361	1396	2289	2615	3749	369	1406	2278<-	361	1396	2615
354	1415	2297	2717	3826	424	1421	2294<-	354	1415	2717
338	1417	2201	2927	0	479	1424	2239<-	338	1417	2927
483	1576	2110	2409	3485	493	1424	2184<-	483	1576	0
482	1453	1998	2395	3657	487	1424	2129<-	482	1453	0
495	1404	1962	2537	3767	530	1424	2068<-	495	1404	2537

Figure 4-7
Example of F3 Values for /en/ in the Hand-edited and the System Files. The Expected Formant Frequency Estimate for a Male Speaker of /en/ is 3159.

The phonemes /ax/, /axr/, /eh/, /ih/, /ix/, /o/, /ux/, /uw/, /uh/, /ow/, /r/, and /w/ were labeled correctly over 90% of the time for each of the three formants. The phoneme /uh/, which had the best results, was labeled 100% for F1, 99.2% for F2, and 96.9% for F3. Phonemes not labeled 85% of the time for any formant are /el/, /en/, /ey/, /iy/, /oy/, /y/. Two of the three syllabic resonants fall into this category. The third (/em/) would probably also, but there were no instances of it in the test database.

Phoneme	F1	F2	F3
AA	99.5%	99.5%	89.3%
AE	98.1%	87.3%	87.0%
AW	99.6%	94.1%	84.1%
AX	98.3%	96.6%	92.2%
AXR	98.3%	92.4%	91.1%
AY	99.9%	90.9%	72.2%
E	NO DATA	NO DATA	NO DATA
EH	99.8%	92.9%	90.8%
EL	97.7%	98.8%	81.7%
ER	98.6%	89.0%	90.5%
EM	NO DATA	NO DATA	NO DATA
EN	100%	94.7%	61.4%
EY	99.1%	78.4%	85.8%
IH	99.6%	96.2%	96.2%
IX	97.9%	94.8%	97.4%
IY	98.9%	83.2%	87.3%
AO	99.6%	99.1%	89.4%
O	99.0%	100%	100%
OE	100%	100%	88.4%
OY	100%	99.1%	81.0%
AH	97%	97%	88.6%
UX	100%	91.8%	100%
UW	99.3%	98.2%	93.9%
UH	100%	99.2%	96.9%
OW	99.7%	99.2%	91.5%
L	98.6%	96.5%	87.9%
R	100%	93.5%	92.6%
Y	99.7%	86.9%	78.7%
W	95.0%	93.9%	93.7%

Table 4-5
Results for Individual Phonemes

There is a wide variation between individual speakers. Even within speakers of a particular sex (for example males), there is a wide variation for F3: MDC with 94.8% correctly labeled versus MJG with 76.2% correctly labeled.

	F1	F2	F3
MDC	99.9%	97.5%	95.1%
MEC	99.1%	97.6%	86.9%
MJG	99.1%	92.4%	76.1%
MJM	98.2%	96.3%	84.2%
FDM	99.8%	90.0%	90.5%
FHG	98.5%	94.9%	95.1%
FJM	98.3%	88.3%	90.7%
FPR	98.9%	88.9%	88.4%

Table 4-6
Results for Individual Speakers

CHAPTER 5

CONCLUSIONS

Every study must be compared against some data which is determined to be correct. This study uses the hand-edited files described in Chapter 3. These files took a long time to create and, for the most part, are accurate. However, a few problems exist with the hand-edited version. Occasionally the formant values for two adjacent formants in one frame are very close in value, but one frequency value is out of line with the other frames in this formant (see Figure 5-1). This appears to be a bug in either the hand-editing or in the LPC Tool.

Values in LPC File					Hand-edited File			System File		
535	1234	1577	2225	3730	0	1371	2419	535	1577	2225
518	1535	2147	2219	3690	546	1627	1695<-	518	1535	2147
524	1520	2159	2304	3687	556	1847	1904	524	1520	2159
504	1525	2161	2625	3715	556	1947	2560	504	1525	2161
483	1552	2136	2663	3720	556	2008	2590	483	1552	2136

Figure 5-1
Error in Software of Hand-editing Program

It is sometimes difficult, even when using all the visual cues available to the expert, to determine where a formant is, or which peaks are the peaks in the formant. Part of this can be attributed to the relatively modest quality of the spectrogram displayed on the monitor, part to error on the part of the expert, and part to occasionally poor quality of speech. Some of the phonemes were edited incorrectly. A common error was the tracing of /r/ as can be seen in Figure 5-2. For

the phoneme /r/, F3 is usually low and very near F2. What was labeled as F3 is probably F4. F3 is within the area of the dotted box. This error occurred more often where the adjacent phoneme to /r/ was a phoneme like /iy/ with a high expected F3 value.

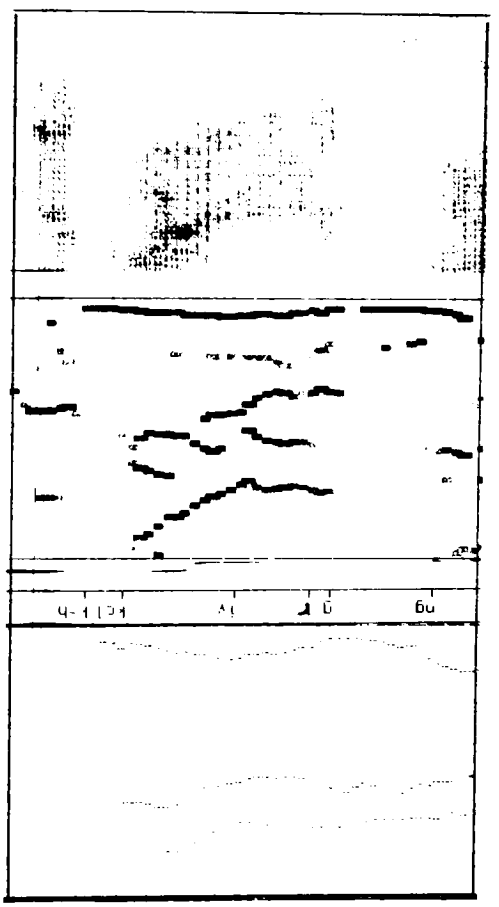


Figure 5-2
Error in Tracing the Phoneme /r/

The estimated expected formant frequency values were gathered from several sources. The values from the Peterson and Barney study [PETE52] are probably the most accurate, and they offer values for female speakers, as well as maximum and minimum expected formant frequency values. Although the phonemes in the

Peterson and Barney study represent fewer than half the phonemes, these are the most frequently used phonemes, therefore representing more than half the instances of phonemes in the study. The maximum and minimum values, as well as all values for female speakers, were derived based on information from [PETE52] and [FANT73]. The poor results for F3 in the syllabics may be attributed to a poor estimate for those phonemes or to antiresonance.

There have been several approaches to tracking formants. Gayvert [GAYV89] used a statistical method in his FSD tracker. Kopec [KOPE86] used a Hidden Markov Model (S40(64) and S40(1024)), also a statistical approach, where 40 refers to the number of divisions in frequency (i.e. to cover 4,000 Hz, each division was 100 Hz) and where 64 and 1024 refer to the number of vector quantizations available for comparison against the target. Markel used peaks from Linear Prediction Coding as did this study. The statistical approaches require a training data set, whereas peak picking using Linear Prediction Coding requires setting of arbitrary thresholds. The objective of the three other studies is to track formants without knowledge of the identity of the phoneme, whereas the objective of the current study was to label the formants using information contained in the phonetic transcription. The RMS results and Percentage of Large Error for the four studies are shown in Figure 5-3. A comparison with McCandless [MCCA70] would have been interesting, but no concrete results were mentioned in the paper. It is assumed that this system performed better because she had no knowledge of the phonetic transcription and simply made an educated guess as to where the peaks should be located, whereas this study had the expected formant frequency values from Peterson and Barney, et. al.

	RMS Error			% Large Error		
	F1	F2	F3	F1	F2	F3
Current Study	90	199	294	1.8	3.5	7.5
Markel	92	274	503	1.2	5.4	21.3
S40(1024)	72	93	150	1.6	0.4	1.5
S40(64)	98	155	235	3.2	1.3	4.4
FSD	68	122	214	0.7	0.4	3.9

Figure 5-3
Results of Current Study, Markel[MARK72], FSD by Gayvert[GAYV89], and S_{40}(1024) and S_{40}(64) by Kopec [KOPE86]

At the start the objective of this thesis was to write an expert system to label formants using the phonetic transcription, knowledge about individual phonemes, and their coarticulatory effects on other phonemes. There are several reasons why this approach was not feasible. The wide variability between speakers is probably the most difficult problem to solve. There is almost a 20% difference between speaker MDC (95.1%) and speaker MJG (76.1%) in the number of frames correctly labeled for F3. The combinations of phonemes (even disregarding most of the consonants, as this study did) is so numerous, and that, combined with the large amount of data which must be studied for a single phoneme, makes any in depth study of individual combinations overwhelming. As Fujimura [FUJI62] stated regarding nasals, the relative levels vary even within one nasal murmur. This variability often carries over to phonemes adjacent to nasals. It is impossible to write an expert system when the results are unpredictable.

Several improvements could be made to the system. Examining the relationship of categories of phonemes versus individual phonemes would decrease the combination of phonemes to be studied. The amplitudes corresponding to the formant frequency values were not used in the current study and contain some

information that could help to determine which peak is a formant. The data structures are already in place to use amplitudes. All references in the code to amplitudes were deleted because the data used for the hand-edited, and therefore for the current system, had invalid amplitude values. The Match Routine could be improved in a couple ways. It currently uses information from the previous phoneme and the target phoneme. The phoneme window could be expanded to four phonemes: preprevious, previous, target, and next. Matching would occur between the preprevious and previous. That would allow for examination of the boundary values between the previous and the target. The Match Routine examines F1, F2, and F3; then it reexamines each formant. Results might be improved by examining all three formant frequency values at one time along with the corresponding amplitudes.

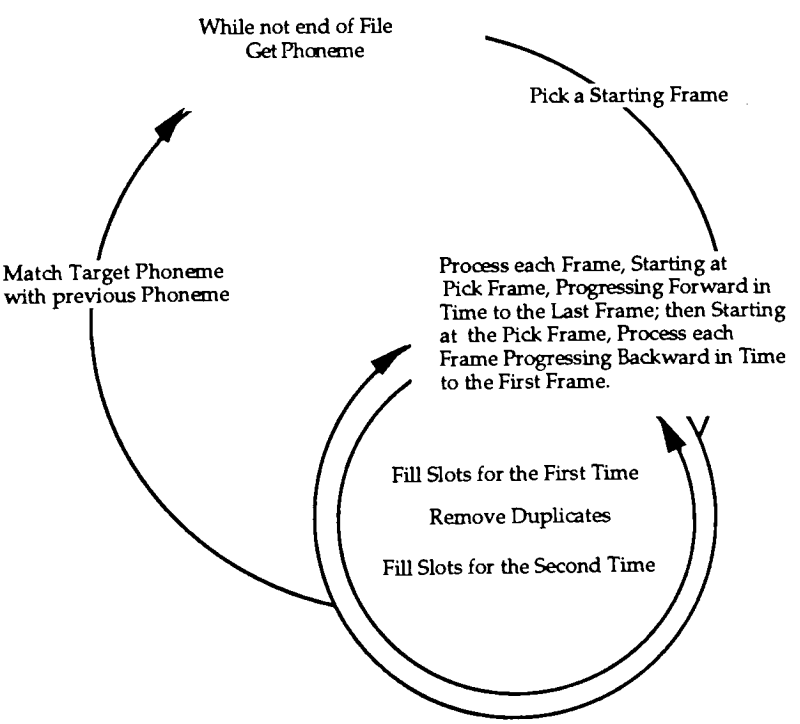
CHAPTER 6

USER DOCUMENTATION

6.1 Description of Project

The purpose of this project is to label peaks in an LPC file as F1, F2, and F3 using the directory name where the files are stored to determine the sex of the speaker, and the phonetic transcription to extract the phoneme and its starting frame.

6.2 Program Flow



6.3 User Inputs

The input to the script file is a directory name. By convention the first letter of a directory is either 'f' or 'm', specifying the sex of the speaker, followed by two initials of the speaker. It is mandatory for the correct operation of the system to have the first letter specify the sex. What follows the first letter is irrelevant to the system as long as it creates a unique directory name in the current directory. For each file within the directory with a .lpc extension, the script will run the program:

mk file.lpc file.l file.r sex

where mk is the name of the makefile; file.lpc is the file containing the LPC peaks and corresponding amplitudes; file.l is the label file containing the phonetic transcription, the starting, and the ending frame for each phoneme to be processed; file.r is the results containing a list of peaks selected as the first three formants; and sex is either the character 'f' or 'm'.

6.4 Discussion of Each Program

6.4.1 Diphthongs: *diph()*

Input: Phoneme Window which is an array of struct phtype.

Output: Multidimensional array 'lpcslot' of struct pktype filled with peaks selected as F1, F2, and F3 for each frame in the phoneme.

Diph() processes all diphthongs as two separate phonemes by dividing each diphthong into 2 parts: the first two-thirds and the last two-thirds, where the center third is common to both parts. The routine fill_slots() is called to fill slots for F1, F2, and F3 with peaks for each frame. If the selection of peaks in the center third is not

the same for the front and back parts, a frame in either the first or second part is used as an estimate and the slots of one boundary third are refilled using fill_slots().

6.4.2 Estimate: *get_estimate()*

Input: A record of struct phtype describing one phoneme.

Output: Field est[3] (est1[3]) filled with expected formant frequencies.

Estimate looks up the expected formant frequencies for F1, F2, and F3 for the specified phoneme. If the phoneme is a diphthong, two values for each of the formants (est[] and est1[]) are returned.

6.4.3 Fill Slots: *fill_slots()*

Input: First_frame, a flag, which indicates if this is the first frame to be processed in the phoneme.

peak[5], an array containing the first five peaks in the frame

est[3], expected formant frequency estimates for the first three formants

Output: slot[3], an array containing the peaks selected as the first three formants

Fill Slots calls

fill_slot1() which fills each (3) slot with a peak that is closest to its corresponding estimate;

remove_duplicates() which, if two or more slots contain the same peak, removes the peak from the slot whose absolute difference between the estimate and peak is larger;

fill_slot_2() fills any empty slots with unused peaks providing the peak(s) meet specific criteria.

6.4.4 Limits: *flimit() & mlimit()*

Input: A record of struct phtype describing one phoneme.

Output: Max[4] (Max1[4]) & Min[4] (Min1[4])

Limits extracts the maximum and minimum expected formant frequency values for F1, F2, and F3 for a particular phoneme. If the phoneme is a diphthong, two values for each formant are returned.

6.4.5 Match: *match_formants()*

Input: Phoneme Window which is an array of struct phtype.

Output: Multidimensional array 'lpcslot' of struct pktype filled with peaks selected as F1, F2, and F3 for each frame in either the previous or target phoneme.

The match routine determines if the boundaries between two phonemes (the previous and the target) are continuous and, if not, examines the unused peaks to select a phoneme (either the previous or the target) whose selected peaks are used as estimates to select peaks in the other phoneme.

6.4.6 Pick Frame: *pick_start_frame()*

Input: Structure of phtype describing one phoneme

Output: Frame (integer)

Pick_start_frame() calls

initialize_array() to initialize an array which will contain point values for each frame;

assign_limit_pt() which assigns each frame points for each peak that is within the maximum and minimum values for that phoneme;

assign_loc_pt() which assigns each frame points for its location within the phoneme;

return_best_frame() which examines the array containing the point values and returns a frame number.

6.4.7 Sort Phoneme: *sortphon()*

Input: the phonetic transcription

Output: an integer denoting the group to which the phoneme belongs

Sort Phoneme is a simple look up.

6.4.8 Tools

Tools consists of the following routines used by multiple programs:

help() prints the syntax for command line if the incorrect number of parameters is entered.

string_to_int() converts characters to an integer.
Input: characater string
Output: integer value of string

diff() returns the absolute difference for the two parameters passed to it.
Input: two integers
Output: integer containing absolute difference

check_dash() takes the phonetic transcription and removes the dash, if one exists, and any characters following the dash.
Input: phonetic transcription ts[8]
Output: phonetic transcription ts[8] with dash and following characters removed

initialize_2d_array() initializes a two dimensional array to zeros.
Input: two dimensional array of integer
row count (int)
column count (int)

init_2d_array_struct() initializes a two dimensional array of struct pktype to zeros.

Input: two dimensional array of struct pktype
row count (integer)
column count (integer)

Output: two dimensional array of struct phktype set to zeros

initialize_array() initializes a one dimensional array.

Input: one dimensional array of integers
length (integer)

storest() stores the current estimates in a temporary array (tmpest[])

Input: est[]
flag set to zero for vowels, semivowels and the front part of a diphthong; set to one for the last part of a diphthong.

Output: tmpest[]

change_est() changes the estimate back to the estimate selected from the database after processing from the pick frame to the end of the last frame.

Input: Structure of phtype describing one phoneme
frame (int)
tmp (int) containing a selected value for F4

Output: est[] containing values from the database

proc_part2() processes each frame from the pick frame to the last frame. It calls fill_slots().

Input: Structure of phtype describing one phoneme
Start Frame (integer)
Estimate[]

Output: Lpcslot[][] from pick frame to last frame filled with selected peaks.

proc_part1() processes each frame from the pick frame to the first frame. It calls fill_slots().

Input: Structure of phtype describing one phoneme
Start Frame (integer)
Estimate[]

Output: Lpcslot[][] from pick frame to first frame filled with selected peaks.

transfer_est() transfers the estimate in the global variable est[] to the field est[] in the struct phtype.

Input: Structure of phtype describing one phoneme

Flag to determine if referring to the last part of a diphthong
Output: Field est[] in struct phtype with values of global est[]

transfer_array() transfers a two dimensional array from one array to another array.
Input: Old array [] []
Row count (integer)
Column count (integer)
Output: New array [] [] containing values of old array [] []

transfer_array_struct() transfers a two dimensional array of struct pktype from one array to another array.
Input: Old array [] []
Row count (integer)
Column count (integer)
Output: New array [] [] containing values of old array [] []

6.4.9 Vowels: *process_vowel()*

Input: the phoneme window, which is an array of struct phtype
the sex of the speaker

Output: the array lpcslot, which is part of struct phtype, for the target phoneme containing the peaks selected as the first three formants

Process_vowel() calls

flimit() or mlimit(), depending on the sex of the speaker, to get the maximum and minimum values;

coart_adjust_vow() to make adjustments to the est[], max[], and min[] based on the adjacent phonemes;

pick_start_frame() to determine the best starting frame;

storest() to store the database estimate to use to start processing the second part from the pick frame;

proc_part2() fills the slots with peaks from the start frame to the last frame;

change_est() changes the estimate, which was altered in proc_part2, to the database estimate;

proc_part1() which fills the slots with peaks from the start frame - 1 to the first frame.

REFERENCES

[ALLE87]
Allen, J., Honnicutt, M. S. T., and Klatt, D., *From Text to Speech: MITalk System*, Cambridge University Press, Cambridge, England, 1987.

[FAIR61]
Fairbanks, G., and Grubb, P., "A Psychophysical Investigation of Vowel Formants," *Journal of Speech and Hearing Research*, Vol. 4, pp. 203 - 219.

[FANT73]
Fant, G., *Speech Sounds and Features*, MIT Press, Cambridge, 1973.

[FLAN56]
Flanagan, J. L., "Automatic Extraction of Formant Frequencies from Continuous Speech," *Journal of the Acoustical Society of America*, Vol. 28, pp. 110 - 118.

[FLAN72]
Flanagan, J. L., *Speech Analysis, Synthesis, and Perception*, Springer Verlag, Berlin, 1972.

[FRY79]
Fry, D. B., *The Physics of Speech*, Cambridge University Press, Cambridge, 1979.

[FUJI62]
Fujimura, O., "Analysis of Nasal Consonants," *Journal of the Acoustical Society of America*, Vol. 34, pp. 1865 - 1875.

[GAYV87]
Gayvert, R., T. and Hillenbrand, J., "Formant Tracking Using Statistical Pattern Recognition," *Journal of the Acoustical Society of America,* Suppl. 1, Vol. 82, S37.

[GAYV88]
Gayvert, R. T. and Hillenbrand, J., "Statisitical Approaches to Formant Tracking," *Journal of the Acoustical Society of America,* Suppl. 1, Vol. 84, S22.

[GAYV89]
Gayvert, R. T., *A Statistical Approach to Formant Tracking,* Masters Thesis Rochester Institute of Technology, 1989.

[HIER86]
Hieronymus, J. L., and Majurski, W. J., "Compensating for Vowel Coarticulation in Continuous Speech Recognition," *ICSSP86,* pp 2787 - 2790, 1986.

[HILL77]
Hillenbrand, J., "The Categorization of Speech Sounds by Infants," Dissertation Proposal, Summer 1977.

[HILL87]
Hillenbrand, J., and Gayvert, R. T., "Speaker-independent Vowel Classification Based on Fundamental Frequency and Formant Frequency," *Journal of the Acoustical Society of America,* Suppl. 1, Vol. 81, S45.

[HOLB62]
Holbrook, A. and Fairbanks, G., ""Diphthong Formants and their Movements," *Journal of Speech and Hearing Research,* Vol. 5, No. 1, pp. 38 - 58.

[KOPE86]
Kopec, G., "Formant Tracking Using Hidden Markov Models," *IEEE International Conference of Acoustics, Speech, and Signal Processing*, Tampa, Florida, pp. 1113 - 1116.

[KUWA85]
Kuwabara, H., "An Approach to Normalization of Coarticulation Effects for Vowels in Connected Speech," *Journal of the Acoustical Society of America*, Vol. 77, pp. 686 - 694.

[LASS76]
Lass, N. J., *Contemporary Issues in Experimental Phonetics*, Academic Press, New York, 1976.

[LASS82]
Lass, N. J., *Speech, Language, and Hearing*, Saunders, Philadelphia, 1982.

[LEHI61]
Lehiste, I., and Peterson, G. E., "Transitions, Glides, and Diphthongs," *The Journal of the Acoustical Society of America*, Vol. 33, No. 3, pp. 268 - 277.

[LEHI64]
Lehiste, I., "Acoustical Characteristics of Selected English Consonants", *International Journal of American Linguistics*, Vol. 30, pp. 1-97.

[LIEB77]
Lieberman, P., *Speech Physiology and Acoustic Phonetics*, MacMillan Publishing Co., Inc., New York, 1977.

[MARK72]
Markel, J. D., "Digital Inverse Filtering--A New Tool for Formant Trajectory Estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-20, pp. 129 - 137.

[MCCA74]
McCandless, S. S., "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-22, pp. 135 - 141.

[MINI73]
Minifie, F. D., Hixon, T., and Williams, F. (eds.), *Normal Aspects of Speech, Hearing, and Language*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1973.

[MON83]
Monsen, R. B. and Engebretson, A. M., "The Accuracy of Formant Frequency Measurements: A Comparison of Spectrographic Analysis and Linear Prediction," *Journal of Speech and Hearing Research*, Vol. 26, pp. 88-97.

[OCON57]
O'Connor, J. D., Gerstman, L. J., and Liberman, A. M., "Acoustic Cues for the Perception of Initial /w, j, r, l/ in English," *Word*, Vol. 13, pp. 24 - 43.

[OHMA65]
Ohman, S. E. G., "Coarticulation in VCV Utterances: Spectrographic Measurements," *Journal of the Acoustical Society of America*, Vol. 39 No. 1, pp. 151 - 168.

[OLIV71]
Olive, J. P., "Automatic Formant Tracking by a Newton-Raphson Technique," *Journal of the Acoustical Society of America*, Vol. 50, pp. 661 - 670.

[PARS86]
Parsons, T., *Voice and Speech Processing*, McGraw-Hill, Inc., New York, 1986.

[PETE52]
Peterson, G. E., and Barney, H. L., "Control Methods Used in a Study of Vowels," *The Journal of the Acoustical Society of America*, Vol. 24, No. 2, pp. 175 - 184.

[SCHA70]
Schafer, R. W., and Rabiner, L. R., "System for Automatic Formant Analysis of Voiced Speech," *Journal of the Acoustical Society of America*, Vol. 47, pp. 637 - 648.

[STEVE63]
Stevens, K. N. and House, A. S., "Perturbation of Vowel Articulations by Consonantal Context: an Acoustical Study," *Journal of Speech and Hearing Research*, Vol. 6, No. 2, pp. 111 - 128.

[STEVE66]
Stevens, K. N., House, A. S., and Paul, A., "Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation," *Journal of the Acoustical Society of America*, Vol. 40, pp. 123 - 132.

[ZUE85]
Zue, V. W., "The Use of Speech Knowledge in Automatic Speech Recognition," *Proceedings of IEEE*, Vol. 73, No. 11, pp. 1602 - 1615.

Appendix A

Carnegie-Mellon University Phonetic Transcription (C-MU)

and

International Phonetic Alphabet (IPA)

C-MU	IPA	Example
I. Vowels		
iy	i	'beat'
ih	I	'bit'
eh	ɛ	'bet'
ae	æ	'bat'
ux	ju	high, front, rounded allophone of /uw/ as in 'beauty'
oe	oe	mid-low, front, rounded allophone of /ow/
ix	ɚ	high, central vowel (unstressed), as in 'roses'
ax	ə	mid, central vowel (unstressed), as in 'the'
ah	ʌ	mid, central vowel (stressed), as in 'butt'
uw	u	'boot'

uh	U	'book'
ao	ɔ	'bought'
aa	ɑ	'cot'
ey	e	'bait'
ay	aI	'bite'
oy	ɔI	'boy'
aw	aU	'bough'
ow	o	'boat'
e	e	nondiphthongized /ey/; c.f. 'eh' (meaning 'What did you say?')
o	o	mid-low, back, nondiphthongized allophone of /ow/

II. Liquids

l	l	'led'
r	r	'red'

III. Glides

y	j	'yet'
w	w	'wet'

IV. Syllabic Resonants

er	ɝ	'bird'
axr	ɚ	unstressed allophone of /er/, as in 'diner'
el	l	syllabic allophone of /l/, as in 'bottle'
em	m	syllabic allophone of /m/, as in 'yes 'em' (meaning yes ma'am')
en	n	syllabic allophone of /n/, as in 'button'
eng		syllabic allophone of /ng/, as in 'Washington'

V. Stops

p	p	'pop'
b	b	'bob'
t	t	'tot'
d	d	'dad'
k	k	'kick'
g	g	'gag'
m	m	'mom'
n	n	'non'
ng	ŋ	'sing'
q	ʔ	glottal stop -- allophone of /t/, as in Atlanta, where the first /t/ can be realized as [q]. Also may occur between words in continuous speech, especially at vowel-vowel boundaries,

		and at the beginning of vowel-initial utterances

VI. Affricates

ch	ʧ	'church'
jh	ʤ	'judge'

VII. Fricatives

f	f	'fief'
v	v	'very'
th		'thief'
dh	θ	'they'
s	s	'sis'
z	z	'zoo'
sh	ʃ	'shoe'
zh	ʒ	'measure'
hh	h	'hay'
hv		'voices allophone of [hh], occurs between vowels

VIII. Flaps and Trills

dx	alveolar flap (allophone of [t] & [d])
nx	nasal flap (allophone of [n])
lx	lateral flap (allophone of [l])
rx	trill (allophone of [r]) -- not used now

IX. Nonspeech

bg	silence at the beginning and end of an utterance
pau	silence within an utterance that does not correspond to the closure for a stop or affricate; usually audible at sentence level
sil	same as [pau], but shorter and not audible at sentence level
ns	a nonspeech sound
h#	exhalation at end of utterance
#h	exhalation at beginning of utterance

X. Other

voi	voicing not associated with a stop closure
cl	at this level, we identify the closure with its stop (e.g. [pcl] means that the closure is for a [p], whether the [p] is released or not)
epi	closure resulting from coarticulation of fricative and nasal or lateral
-h	appended to stops to signify aspiration; appended to any voiced segment to signify devoicing
-n	appended to sonorant segments to signify nasalization
-q	appended to sonorant segments to signify glottalization/laryngealization
-b	appended to stops to signify the release of a stop in an environment where stops are often not released (e.g. [k-b] as in 'black board' if the [k] in 'black' is released or in clause-final position)

Appendix B

Estimates, their Corresponding Maximum and Minimum Expected Frequency Values, and the Source of this Values

Males

Phoneme	F1	Min1	Max1	F2	Min2	Max2	F3	Min3	Max3	Source
iy	267	190	420	2297	2000	2700	2800	2635	3590	P & B
ih	393	206	525	2010	1750	2300	2599	2340	3100	P & B
eh	525	370	640	1860	1670	2140	2502	2260	2860	P & B
ae	662	514	830	1727	1470	2140	2441	2110	2970	P & B
oe	341	186	497	1581	1294	1939	2308	1944	2785	C-MU
ix	420	344	534	1680	1446	1936	2520	2076	3000	MITTalk
ax	550	474	664	1260	1026	1516	2470	2088	2973	MITTalk
ah	626	550	740	1194	960	1450	2394	1950	3000	P & B
uw	439	310	562	1036	730	1642	2298	1860	3500	P & B
uh	304	210	400	895	650	1320	2267	1850	3300	P & B
ao	573	430	720	845	550	1050	2462	1950	3480	P & B
aa	719	550	840	1096	760	1330	2484	2020	2930	P & B
e	475	300	600	1900	1770	2240	2502	2260	2860	derived
o	489	410	662	1096	830	1742	2298	1860	3500	derived
l	330	270	640	1050	870	1740	2800	2460	3160	MITTalk
r	330	250	640	1060	760	1330	1380	1020	1930	MITTalk
y	240	170	540	2070	1260	2450	3020	2450	3300	MITTalk
w	285	230	520	610	450	950	2150	1350	2880	MITTalk
er	486	360	590	1358	1130	1561	1733	1400	2800	P & B
axr	520	343	709	1400	1132	1729	1650	1356	2020	MITTalk
el	450	284	624	800	596	1031	2850	2427	3414	MITTalk
em	200	61	334	900	686	1147	2100	1758	2542	MITTalk
en	200	61	334	1600	1311	1961	2700	1758	2543	MITTalk
eng										
ux	290	141	438	1900	1579	2310	2600	2315	3124	MITTalk
	330	177	485	1200	954	1496	2100	1758	2543	MITTalk
ey	480	310	659	1720	1418	2100	2520	2133	3031	MITTalk
	330	177	485	2200	1846	2660	2600	2204	3124	MITTalk
ay	660	471	868	1200	954	1496	2550	2160	3066	MITTalk
	400	239	566	1880	1561	2287	2500	2115	3007	MITTalk
oy	550	373	740	960	740	1217	2400	2026	2892	MITTalk
	360	203	520	1820	1508	2217	2450	2070	2950	MITTalk
aw	640	454	845	1230	981	1531	2550	2160	3066	MITTalk
	420	257	589	940	722	1194	2350	1981	2834	MITTalk
ow	540	363	729	1100	865	1380	2300	1936	2775	MITTalk
	450	284	624	900	686	1148	2300	1936	2775	MITTalk

Females

Phoneme	F1	Min1	Max1	F2	Min2	Max2	F3	Min3	Max3	Source
iy	320	200	450	2792	2360	3100	3129	2945	3840	P & B
ih	450	319	535	2488	2090	2780	3070	2740	3400	P & B
eh	611	418	760	2355	1980	2570	3020	2700	3500	P & B
ae	864	650	1110	2064	1700	2560	2849	2300	3260	P & B
oe	400	239	560	1850	1534	2252	2700	2294	3240	C-MU*
ix	491	320	666	1880	1627	2155	2870	2482	3352	MITTalk*
ax	675	535	836	1460	1000	1735	2770	2380	3626	MITTalk*
ah	759	618	920	1413	1160	1688	2768	2380	3250	P & B
uw	478	315	690	1171	860	1506	2707	2300	3240	P & B
uh	380	290	480	977	630	1430	2668	2260	3100	P & B
ao	589	380	720	925	590	1110	2739	2250	3300	P & B
aa	863	592	1030	1226	1020	1470	2780	2290	3180	P & B
e	571	360	700	2395	2200	2670	3020	2700	3500	derived
o	520	415	790	1201	960	1606	2707	2300	3240	derived
l	398	270	640	1265	870	1740	3000	2562	3589	MITTalk*
r	412	250	640	1240	989	1643	1725	1020	1930	MITTalk*
y	289	170	540	2422	2045	2917	3300	2450	3500	MITTalk*
w	343	230	520	735	450	950	2590	1350	2800	MITTalk*
er	512	360	652	1649	1240	2120	1981	1440	2480	P & B
axr	627	442	830	1687	1389	2062	1988	1658	2412	MITTalk*
el	527	284	624	936	596	1031	3200	2427	3414	MITTalk*
em	234	61	334	1053	822	1325	2457	2294	3240	MITTalk*
en	234	61	334	1872	1554	2278	3159	2700	3500	MITTalk*
eng										
ux	340	186	496	2223	1868	2686	3132	2680	3742	MITTalk*
	398	237	564	1446	1174	1782	2530	2142	3042	MITTalk*
ey	578	398	773	2072	1733	2510	3036	2594	3631	MITTalk*
	398	237	564	2650	2249	3182	3132	2680	3743	MITTalk*
ay	795	592	1025	1446	1174	1782	3072	2626	3673	MITTalk*
	482	312	661	2265	1905	2735	3012	2575	3603	MITTalk*
oy	663	474	871	1157	915	1446	2892	2465	3463	MITTalk*
	434	270	605	2193	1840	2650	2952	2519	3530	MITTalk*
aw	771	571	997	1482	1205	1824	3072	2626	3673	MITTalk*
	506	334	689	1132	893	1417	2831	2410	3393	MITTalk*
ow	650	462	857	1325	1065	1642	2771	2357	3323	MITTalk*
	542	366	731	1084	850	1361	2771	2357	3323	MITTalk*

GLOSSARY

allophone — a phoneme which when produced in different contexts or repeated at different intervals is not identical to the original phoneme, but is still perceived as a member of that phoneme

alveolar ridge — an area of the hard palate between the cuspids and the upper molars

amplitude — the range or extent of movement of a vibrating body

antiresonance (zeros) — an antiresonance eliminates a resonance tuned to the same frequency and reduces the effect of an adjacent resonance

articulation — the movements of the structures above the larynx that serve to change the configuration of those cavities.

articulation (manner of) — the source of sound generation such as plosive (burst), fricative (turbulence), or sonorant and vowel (glottal vibrations)

articulations (place of) — place in the vocal tract where the constructed airway generates turbulence

C-MU notation	an ASCII character representation of the phonetic alphabet used at Carnegie-Mellon University (see Appendix A) .
CVC	consonant vowel consonant string
centralization	the phenomenon where formant frequency values tend to fall in a more central location than expected values because of the effects of coarticulation
coarticulation	the merging or overlapping of two adjacent sounds because the tongue, lips, jaws, etc. cannot jump from one target position to the next, but instead performs a blending process which gives rise to a smooth and continuous flow
consonant	a voiced or voiceless speech sound which may be used to link vowels together
damping	the decreasing of amplitude of vibrations of a sounding body caused by the absorption of energy by the surrounding medium
diphthong	two vowels that are blended together within one syllable
formant	a resonance of the vocal tract
formant (first)	the lowest frequency peak of vocal tract energy, which decreases as the height of the tongue increases

formant (second)	the second lowest frequency peak of vocal tract energy which is high with front vowels and low with back vowels
Fourier analysis	the analysis by which the spectral frequencies composing the complex wave are determined
frequency	the rate of vibration; the number of cycles that occur each second
glottis	the opening between the vocal folds
hard palate	the bony roof of the mouth
Hertz (Hz)	the measure of cycles per second (cps)
International Phonetic Alphabet (IPA)	widely accepted phonetic alphabet containing nonascii characters (see Appendix A).
intensity	the energy released per unit time
labial	pertaining to the lips or sounds produced by the lips
larynx	a speech organ which acts as a valve between the lungs and the mouth. It controls the airflow from the lungs by opening and closing
linear prediction	

low-pass filter	a device, electrical or mechanical, which attenuates high=frequency energy and allows low-frequency energy to pass through
LPC files	those files used by this study that were created by linear prediction coding
nasal	formation of one or more oral closures as air flows through the nose
nasal cavity	the cavity extending from the nasopharynx to nares
nasalized	is a configuration of the vocal tract in which there is an intermediate velopharyngeal opening and in which there is not an oral closure
nasopharynx	the upper throat from the soft palate to the base of the skull
null environment	an environment that does not exert any strong coarticulatory effects on the adjacent phoneme. Specifically in the Peterson and Barney study it is the /h vowel d/ context
pharynx	the common cavity of the respiratory and digestive tract; the throat
phonation	the production of sound by vibration of the vocal cords
phone	an individual speech sound

phoneme	the smallest distinctive group or class of phones in a language; American English has forty-three common phonemes (excluding regional dialects)
phonetics	scientific study of speech sounds that occur in a given language
poles	see resonance
resonator	a cavity that selectively filters out energy at frequencies to which it is not 'tuned' and amplifies at frequencies to which it is 'tuned'
resonance	the phenomenon whereby one body, which has a tendency to vibrate at a certain frequency, will build up vibrations with comparatively large amplitude when it is set into motion by another body which is vibrating at a similar frequency
retroflex	configuration of the tongue in which the apex is turned back so that the blade has a concave shape from above the forsoventral dimension
semivowels	phonemes that have characteristics of both vowels and consonants
soft palate (velum)	the movable portion of the palate that can open or close the nasal cavity from the oral and pharyngeal cavities
sonorant	a voiced sound generated through a constriction that does not cause turbulence

sound spectrograph	electrical equipment which produces visible speech showing the frequency, intensity, and time elements of the sound
spectral envelope	a diagram showing the relative amplitude of the frequency components of a sound by smoothly connecting the peaks of intensity
spectrogram	the output of a spectrograph; a spectrum of frequency and intensity recorded across time
spectrum	as it pertains to sound, a representation of the amplitude of the components arranged as a function of their frequencies
steady state	a continuous segment of speech that does not change in quality
transfer function	the frequency distribution of the filters (resonators) of the vocal tract that shape the vowel and consonant energy
transition	the sound produced by movement of the vocal tract from an occluded to an open position
turbulence	noise generated by friction of air flow through constrictions in the vocal tract
undershoot	the target location of a phoneme is not reached because of the coarticulatory effects of nearby phonemes

unvoiced	sound produced when the glottis is open and not vibrating
velum (soft palate)	see soft palate
vocal folds	(also known as the vocal cords) a paired muscle located in the larynx
vocal tract	consists of the pharyngeal, oral, and nasal cavities
voiced	sound produced when the glottis is closed and vibrating
vowels	a voiced sound produced by relative free passage of the air stream through the larynx and oral cavity
zeros	see antiresonance