Recognition of English Vowels using Top-down Method

by

Park Chi-youn

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2004

Author ....
Department of Electrical Engineering and Computer Science
May 20, 2004

Certified by ....
Kenneth N. Stevens
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Recognition of English Vowels using Top-down Method

by

Park Chi-youn

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2004, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Many recognizers use bottom-up methods for recognizing each phoneme or feature, and use the cues and the context to find the most appropriate words or sentences. But humans recognize words not just through bottom-up processing, but also top-down. In many cases of listening, one can usually predict what will come based on the preceding context, or one can determine what has been pronounced by listening to the following sounds. Therefore, if some cues to a word are given, it would be possible to refine the recognition by using the top-down method.
This thesis deals with the improvement of the performance of recognition by using the top-down method. And most of the work will be concentrated on the problem of vowel recognition, when the adjacent consonants are known.

Thesis Supervisor: Kenneth N. Stevens
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank my thesis supervisor, Kenneth Stevens, for his guidance into the world of speech processing during my first two years at MIT, expecially for his patience and kindness during the periods while I did not show up at the office.

Thanks to many friends at MIT who always supported me through the whole year of working on the thesis. Thanks to many friends in Korea who encouraged me while I was too busy, and while my work did nto go well.

Finally I want to thank my parents who supported and believed me all through the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Levels of Recognition

Recognizing a spoken sentence means choosing one sentence among infinite number of possible sentences that fits the given speech signal the best. But we can not compare a given speech signal with every possible sentence, because there are infinite number of sentences. As comparing every sentence with a given speech signal is not possible, most recognizers break up the sentences into words, which is more restricted in number. But as the number of all the words is still more than 100,000, it is not practical to compare all of them. So they split the words to smaller units of speech, or phonemes. And now they can compare and choose among just dozens of possibilities. Those various levels are depicted schematically in Figure 1-1

Therefore, for most of the speech recognition systems, the phonemes are usually recognized first, and using the information of phonemes, words are recognized, and using the information of words, the sentences are recognized. This way of processing is called a bottom-up processing.

Sentence	I have a book							
Word	I	have			a	book		
Phoneme	a^y	h	ae	v	e	b	u	k

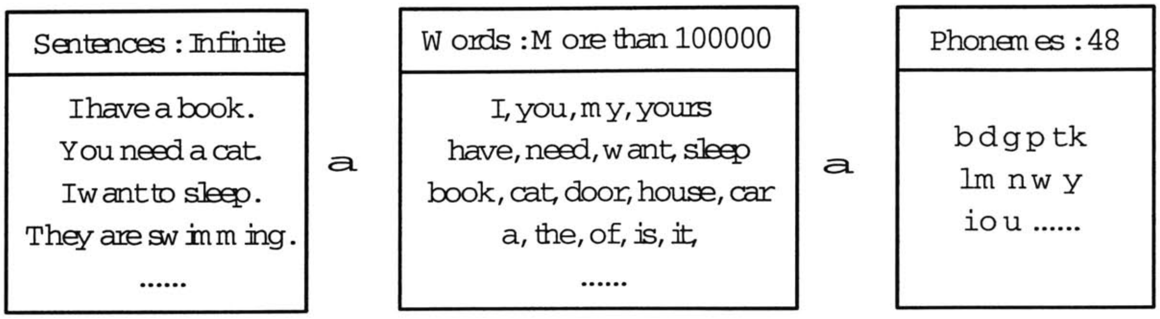

Figure 1-1: Relations between three different layers used in speech recognition(above), and examples and the number of elements in each layer

1.1.2 Influence between Phonemes

(a) Overlap in Time

Performing the recognition from the bottom level (phoneme) to the top (sentence), step by step without any interaction between each level is not quite feasible, because the phonemes are not completely separated in time. For example, look at Figures 1-2.

On the left side of Figure 1-2 is a spectrogram of the word 'book'. In this figure, the vowel part is clearly seen in the interval between 0.12ms and 0.28ms. The consonants /b/ and /k/ are located before and after the vowel respectively. But even if there

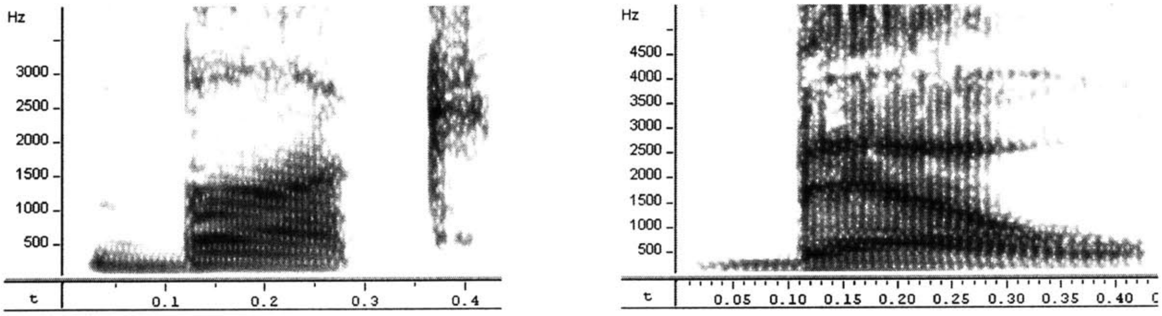

Figure 1-2: Spectrogram of the word 'book'(left) and 'bell'(right)

are clear boundaries between the vowel and consonants, it does not mean that the phonemes are separated clearly. Look at the formants in the vowel interval closely. From 0.2ms, the second formant is rising up significantly and the third formant is getting lower. This means that the vowel is under the influence of the phoneme /k/.

And on the right side of Figure 1-2 is the spectrogram of the word 'bell'. In this case, not only the phonemes /ɛ/ and /l/ affect each other, but the boundary between them is not clear enough to separate the two. Therefore, if we do not consider the interaction between the phonemes, we will have a lot of much difficulty in recognizing them automatically.

(b) Context Effect in Recognition

The problem mentioned above is an interaction of the acoustic consequences of phonemes in the utterance of a word. The same kind of problem arises in hearing a word.

For an example, if one speaks a word 'gift' with an aspirated /g/ so that it sounds like /k/, people are still apt to think that it is a /gɪft/ rather than a /kɪft/. But whe a word 'kiss' is pronounced similar to /gɪs/, people will think the sound of /kɪs/. Such a phenomenon is called the right context effect.[13] Not only does the right context influence the recognition, but also the left context influences the recognition. If one speaks /bʊg/ rather than /bʊk/, people still recognize it as 'book'. This shows that individual phonemes are not recognized separately by a human, but the context around the phoneme may influence the recognition of it.

Because of such interactions between phonemes, we may not be able to separate the phonemes clearly in time, and even if we can manage to separate them somehow, they may not be recognized separately, because of the context effects.

1.1.3 Solutions for the Problem

(a) Probabilistic Method

To deal with these problems, there are several different approaches. One of these approaches is to recognize phonemes not by a deterministic method, but probabilistic.

Therefore, the problem of an incorrect recognition at one level can has less effect at the higher level. By scoring the likelihood of each possible candidate sentence using the context cues such as bigram or lexicon, the effect of error can be compensated, and a better result can be obtained.

(b) Bigram Method

Another approach is to use pairs of phonemes, rather than each distinct phoneme. If the phonemes have to interfere with each other, it would be better to use the interfered signal for recognition. For example, look at the right of Figure 1-2. Recognizing /bɛ/ and /ɛl/ would give a better result than /b/, /ɛ/, /l/, because we may not need to find the boundary between /ɛ/ and /l/, but we may just look at the interaction of /ɛ/ and /l/ without separating them

This may reduce the problem of overlapping, but there is a possibility that /ɛl/ may still be affected by /b/, and /bɛ/ may not be clearly separated from the waveform. But the importance of separation will be reduced a lot.

(c) Using Distinctive Features

The phonemes interact with each other a lot. Then we may adopt a completely different set of speech units, which are not as prone to be influenced by each other. Such a unit is the distinctive feature, of which there are about 15-20. [12]

Features usually represent the articulator movement, such as the position of tongue body (High/Low, Front/Back), the soft palate (Nasal), or the lips (Round). Therefore, if two phonemes have similar sets of features, they can be thought to have similar sounds. And we may be able to use the property to get a better recognition accuracy.

(d) Top-down Method

The approaches listed above are bottom-up. In the bottom-up process, one should complete the step of phoneme level and then move to the higher level using the results of the lower level. This means that if there is any problem or ambiguity in the lower level process, this will be conveyed to the higher level in any rate. This may not be a very satisfactory procedure, since it could be prone to error.

There are several methods that are not just bottom-up. One top-down approach is the TRACE model. [8] It uses three different levels — feature, phoneme, and word. In this method, too, a word is broken up into smaller units, but we do not do the recognition in just one direction from the lower level to the higher one. The three levels interact with each other, and they reinforce or suppress each other's information. This method may give a better result than just bottom-up methods in that an ambiguity in one layer can be corrected with the help of the other layers.

1.2 Purpose of Research

The TRACE model shows an interaction among the levels of feature, phone, and word. And most of the interactions depend on the statistical result. This thesis will be mostly concentrated on using knowledge about the phonemes and features in these interactions between the levels.

If the consonants next to a vowel are known, certain modifications caused by the consonant on the vowel can be expected, So by compensating for the effect of the consonants, the vowel can potentially be recognized better.

The possible effect of consonants are

1. The formant frequencies can be moved higher or lower, depending on the consonants adjacent to the vowel. For example, when /l/ is adjacent to a vowel, that vowel is going to have a higher F3 frequency and lower F2 frequency. Such effects can be reduced if we know the consonantal context.
2. The duration of a vowel can be influenced by the consonant context, too. It is known that a vowel duration is lengthened if a voiced consonant is adjacent, and reduced if a unvoiced consonant is adjacent. Using this clue, we may separate the tense vowels from the lax vowel with more accuracy
3. The number of possible candidates are reduced. If a consonantal context is known, the number of possible vowels that can fit is reduced to about 3–6. Therefore, the recognizer may reduce the number of computation by ignoring the features that are expected to have no importance. And also by doing this,

the accuracy of the recognition can get higher.

4. Better formant tracking can be made. If the consonant is known, the effects of the consonant that affects the tracking of the vowel — nasalization, aspiration, introduction of pole–zero pairs — can be compensated in formant tracking.

In this research, we will try to solve most of the problems listed above to get a better recognition accuracy.

Chapter 2

Recognition Method

2.1 Limited Candidates

If the consonantal context is known, one of the advantages is that there may be a limited number of vowels that can fit to the context. For example when a word is knwon to be b-(vowel)-g, the word can be one of only six words — bag, beg, big, bog, bug, burg. Therefore, all the other vowels do not need to be considered. This may result in a better performance.

(a) Reduction of Computation
This may reduce some of the computation. For example, if we have /æ/ and /ɛ/ as candidates, we only need to consider the duration of the vowel. If it is quite long, it can be considered as /æ/ and if it is short, it may be considered /ɛ/. When the duration test could not give a significant accuracy, the other factors may be considered.

(b) Improvement of Performance
Having a reduced number of candidates can lead to a better performance. For example, if a word is known to be of the form b-(vowel)-g, /big/ or /bug/ cannot be the result, because it is not an utterance of a word. For those cases, the impossible vowels will be discarded, and /bɪg/ or /bʌg/ may be hypothesized instead.

Features	i	ɪ	ɛ^y	ɛ	æ	ɑ	ɔ	ʌ	o	ʊ	u	ɝ	ɑ^y	ɔ^y	ɑ^w
Diphthong			+										+	+	+
Tense	+	-	+	-	+		-	-	+	-	+				
High	+	+	-	-	-	-	-	-	-	+	+				
Back	-	-	-	-	+	+	-	-	-	-	-				
Front	+	+	+	+	+	-	-	-	-	-	-				
Retroflex												+			

Table 2.1: Features of English vowels

2.2 Feature Based Method

In the proposed recognition system, the distinctive features are determined, and the phoneme is considered as a bundle of distinctive features. This, each vowel can be represented as a bundle of features, as in the table 2.1. Using a feature based system is effective in several ways.

(a) Adaptive for Different Candidates

A feature based model can be easily adapted for individual cases. For example, suppose the candidates are reduced to /i/ and /æ/, then we only need to consider if it is a high vowel or a low vowel. Or if we have /ɝ/ and /i/ and /ɪ/, we may first consider if it is a retroflex, and then if it is tense. We may make a different hierarchical tree for recognition for different sets of candidates.

(b) Applying the Consonant Context

The features represent the status of the articulatory system – the position of tongue, tension in the mouth, etc. Therefore these articulatory parameters are likely to be affected by the consonant directly. For an alveolar consonant, F2 tends to be high, and can influence the adjacent vowel. Thus, we may consider the F2 position for the consonant in determining if the vowel is a front vowel or a back vowel.

2.3 Probabilistic Result

The result of the recognition of the vowel should not be deterministic. Rather, the result should indicate the probability of the given signal to be each of the possible candidates. If a vowel is determined in terms of probability, some of the errors may be compensated in other processes. Also, by using the probability, we may think of the confidence of a certain result. A vowel that is determined to have probability of 0.99 is likely to be correct. But a vowel that is determined to have probability of 0.25 may have some error even if the vowel has the highest probability among the possible candidates.

2.4 Limitation of the Method

This recognition system for vowels is built and evaluated on the assumption that we have determined certain details of the consonant environment and the word boundaries, from other kinds of analysis.

(a) Word Boundaries

We should have prior knowledge of the word boundaries to perform this kind of recognition. The consonant context can still affect the vowel when it is in an adjacent word, but we may not be able to reduce the number of candidates if we do not know the word boundaries. This may be another issue for the top-down method that we are investigating here.

The method which is dealt with in this thesis should be used in such a case that we have already reconstructed most of the sentences and need to know, for example, if a word is 'shoot' or 'shot'.

(b) Exact Consonant Recognition

If the consonants were not exactly recognized, we will not be able to apply the method. Recognizing a vowel when we have only some of the features of the adjacent consonants can be studied later to built a better top-down recognition system.

Chapter 3

Analysis

3.1 Creating a Database

In this chapter, a number of consonant–vowel–consonant words will be analyzed, and the best way to determine each of the vowel features is discussed, assuming that the consonants are known.

(a) Vowels in the Database

To do the analysis, a database of CVC(Consonant–Vowel–Consonant) words is needed. The set should consist of different pairs of vowels and consonants. There are many different vowels, including diphthongs. Table 3.1 shows all of the 18 different English vowels.

This thesis will focus on the recognition of monophthongs. The schwas, or re-

/i/	beat	/ɔ/	bought	/ɑ^{y}/	bite
/ɪ/	bit	/ʌ/	but	/ɔ^{y}/	Boyd
/ɛ^{y}/	bait	/o/	boat	/ɑ^{w}/	bout
/ɛ/	bet	/ʊ/	book	/ə/	about
/æ/	bat	/u/	boot	/ɨ/	roses
/ɑ/	Bob	/ɝ/	Bert	/ɚ/	butter

Table 3.1: Table of English vowels. There are 18 vowels in English. Four of them are diphthongs, three are schwas. Remaining eleven monophthongs will be considered in this thesis.

/i/	/ɔ/	/ɪ/	/ʌ/
/o/	/ɛ/	/ʊ/	/æ/
/u/	/ɑ/	/ɝ/	

Table 3.2: The eleven vowels that are included in the database

	Voiceless	Voiced
Fricatives	/f/,/θ/, /s/, /ʃ/	/v/, /ð/, /z/, /ʒ/
Stops	/p/, /t/, /k/	/b/, /d/, /g/
Nasals		/m/, /n/, /ŋ/
Semivowels		/y/, /w/
Liquids		/r/, /l/
Affricates	/ʧ/, /ʤ/	
Aspirant	/h/	

Table 3.3: Table of English consonants.

duced vowels (/ə/, /ɨ/, /ɚ/) will not be considered, because they are quite context-dependent, and also they may be too short to be recognized in detail. So the reduced vowels and the diphthongs are not included in this database. However, all the other eleven strong and weak vowels are included. These vowels are listed in Table 3.2.

(b) Consonants in the Database

There are 24 different consonants in English. All of those 24 consonants are considered. These consonants can be classified as in Table 3.3.

(c) Words in the Database

A complete database would have all the possible pairs of consonants and vowels. However, this is not plausible, since there should be 24 × 11 × 24 = 6336 different CVC triplets. Recording such a great number of words would take too much time, and also analyzing them all would take a great amount of time and would not give a significantly better result. Thus a smaller set of words is selected to do the work efficiently.

It was decided that three different databases would be used. All three databases should include the CVC words, each of which is in itself is a word. And each database has one of the properties listed below.

1. Words with the same consonants pairs are included. (bag, big, bog, ...)
2. Words with the same CV pairs are included. (bag, bat, bang, bear, ...)
3. Words with the same VC pairs are included. (dab, gab, jab, ...)

The list of the words in each database is given in Appendix A. The database was recorded by four different speakers — two male and two female. The utterances were low-pass filtered with a cut-off frequency of 7 kHz and digitized at a sampling frequency of 16 kHz.

3.2 Processing the Utterance Before Analysis

(a) Labelling Vowel Boundary
Before the analysis is performed, several preliminary details are required. First, as this work assumes that the consonants are already recognized, it can be assumed that the boundary between the consonant and the vowel is already determined. Therefore, the landmarks were recorded at the boundaries of each vowel. This process was done manually.

In many cases, there is no distinct boundary between the consonant and the vowel, expecially a vowel and a semivowel, or between a vowel and a liquid. For those cases, the boundary was set to be the place where the voicing ends. These boundaries were chosen in such a manner that the formants can be tracked continuously between the boundaries.

(b) Getting the Formant Tracks
After the vowel boundaries were set, the formant tracks were computed automatically using the method of Viterbi search which is widely used for formant tracking.[14] This was done by the following steps. Both MATLAB and C were used to perform this task faster and easier.

[Using MATLAB]

1. Down-sample each wave file to 10 kHz.

2. Calculate 12th order LPC every 5ms using 256 samples, pre-emphasized by the factor of 0.7, windowed by a 500 sample Hamming window.
3. Solve LPC to get formant candidates and their bandwidths.
4. Save formant candidates to wave.f, and bandwidths to wave.b

[Using C]

1. Read the formant and bandwidth candidates from the files generated by MATLAB.
2. Sort by the formant frequencies and take the 6 largest formants to exclude the negative frequencies.
3. Perform Viterbi search to find the minimum-scored formant tracks.
   (a) Each step is determined by choosing 4 of the 6 candidates.
   (b) Each track candidates are scored by the following

$$50 \times (\text{sum of differences of formants}) + \text{bandwidths}$$

4. Save the formant track into wave.fmt

And for a more exact analysis, the formant was hand corrected. The hand corrected formants will be used only for the analysis; for the recognition, uncorrected ones will be used.

3.3 Duration

The first feature to be considered is duration. The duration of a vowel is affected significantly by the consonants around it, as well as by the tense-lax feature.

To put it simply, the vowel duration is lengthened if it is adjacent to a voiced consonant, and the duration is shortened if it is adjacent to a unvoiced consonant. But in fact, there are various other factors that affect the duration of a vowel. The vowel duration depends largely on the adjacent consonant, but it also depends on the speaker, and the vowel duration is longer when the word is stressed.

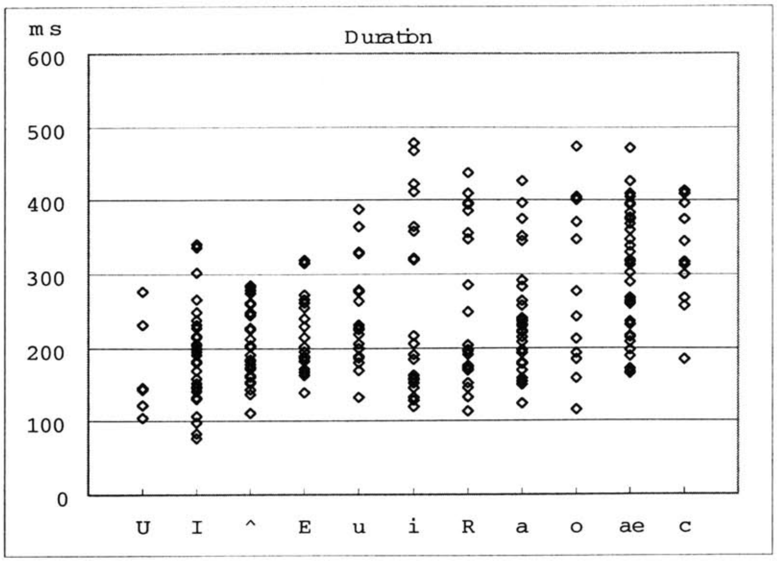

Figure 3-1: The durations of English vowels without conpensating for the consonants. Most of the vowels have large variances.

But in general, the difference between the shortest and the longest vowel is quite large. Therefore, although the vowel duration depends on many different types of environment, at least the vaerage durations of the shortest vowel /ʊ/ and the longest vowel /æ/ are quite different.

As was mentioned in the previous section, it is not possible to set an exact boundary between a vowel and a consonant when the consonant is a liquid or a glide. In those cases, the labels were put at the end of voicing. Therefore, the words that have liquids or glides in it are not analyzed for the duration.

Durations of each of the English vowels for all utterances are plotted in Figure 3-1, and the statistics of these English vowels are given in Table 3.4.

But this does not show a clear separation between vowels. The lax vowels — ʊ,

	ʊ	ɪ	ʌ	ɛ	u	i	ɝ	ɑ	o	æ	ɔ
Average	170	191	210	216	245	246	257	233	295	298	323
SD	68	65	50	54	71	122	108	77	111	84	69

Table 3.4: Averages and standard deviations of the durations of English vowels without compensation. The difference between the average of vowel durations range from 170 to 323 ms.

ɪ, ʌ, ɛ — are usually much shorter than the tense vowels. Also the variances of the lax vowels are much smaller than those of tense vowels. And because the difference between the shortest vowel and the longest vowel is about 150ms, if we can find a way of reducing the variation, it would be very helpful in classifying the tense and lax vowels.

(a) Compensation According to Consonant Context

The consonant effects should be compensated to achieve a lower variance. To compensate for the consonants, we will assume that the existence of a certain consonant context will lengthen the vowel duration by adding a number to an intrinsic duration of the vowel. We also assume that this length correction constant is independent of the vowel. Let us call the constant the lengthening factor of the given consonant. It is necessary, then, to determine the lengthening factor for each consonant. As there is not a sufficient number of words to calculate the accurate rate of change for every vowel, the consonants with similar manner of production are grouped together. The lengthening factor of each consonant is determined so that the sum of the standard deviations of the durations of each vowel is the smallest, i.e., the lengthening factors are adjusted so that

$$\sum SD(v_i)$$

is a minimum, where v_i represents the duration of vowel i.

The lengthening factors that are calculated are given in Table 3.5. We can see that the voiced consonants have positive values, which means that the voiced consonants lengthen the vowel durations. The unvoiced consonants have negative values, as they shortenes the durations. The value of the consonants at the final position are larger than those at the initial position, because the vowel duration is affected more significantly by the consonants following the vowel. The plot of the compensated duration is shown in the Figure 3-2. Table 3.6 shows the averages and standard deviations of the compensated durations.

From Table 3.6, it can be seen that the standard deviations are almost halved, compared to Table 3.4. And still the difference between the longest vowel and the

	b, d, g	p, t, k	m, n, ŋ	f	s	θ	v, z	ʤ	ð	others
Initial	18	-4	20	15	0	15	0	32	78	0
Final	67	-37	73	32	-18	32	136	79	136	0

Table 3.5: Lengthening factors of each vowel. Initial means that the consonant is pronounced before the vowel, and Final means that it is pronounced after the vowel.

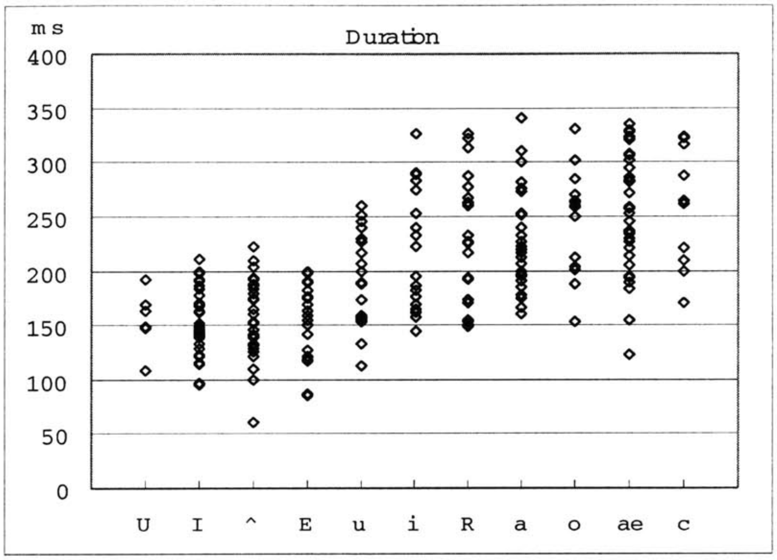

Figure 3-2: The compensated durations of English vowels. The variations are reduced significantly.

	ʊ	ɪ	ʌ	ɛ	u	i	ɝ	ɑ	o	æ	ɔ
Average	154	151	158	152	193	216	228	229	246	255	260
SD	28	29	36	35	42	52	57	46	48	53	55

Table 3.6: Averages and standard deviations of the compensated durations of English vowels

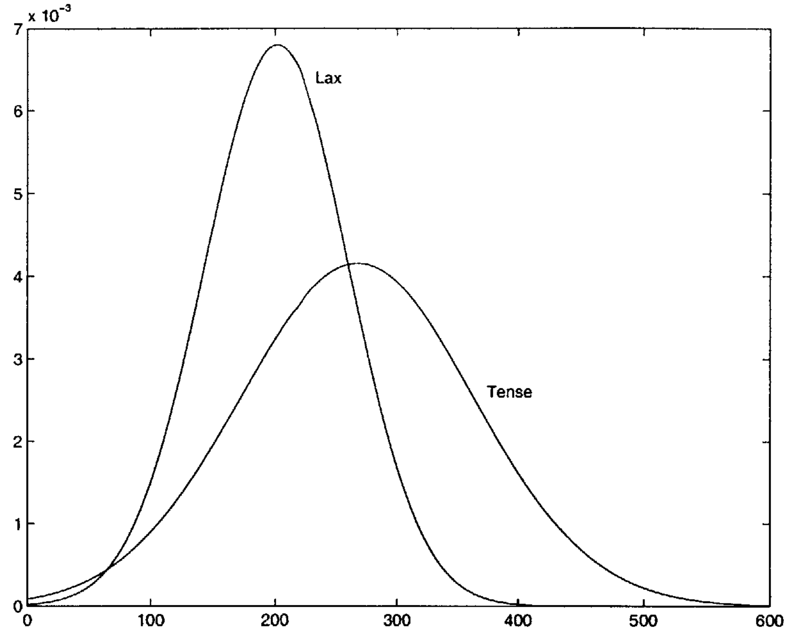

Figure 3-3: The probability of durations of tense and lax vowels before compensation. There are large amount of overlap between the two probability distribution functions.

shortest vowel is about 110ms. Therefore this result can be useful in classifying tense vowels and lax vowels. The probability distributions of the durations of tense and lax vowels are shown in Figure 3-3 (before compensation) and Figure 3-4 (after compensation). Compensation for the consonant gives a better discrimination.

3.4 Retroflex

When pronouncing a retroflexed vowel, or /ɝ/, speakers raise the tongue tip close to the roof of the mouth, so that there is a space underneath the tongue. Such a configuration introduces a pole-zero pair near 2000Hz. This newly-made pole is recognized as F3 formant. Therefore, F3 goes significantly low for a retroflex, even though the exact value of F3 may depend on the speaker or on the context around it. Because of this, a retroflex vowel can be checked by looking at the F3 value.

But the method of analyzing F3 frequency may not be effective for a vowel that is adjacent to the /r/ sound. Of course, there is no syllable in which retroflex /ɝ/ is pronounced before a consonant /r/, because /ɝ/ can also be thought as schwa or

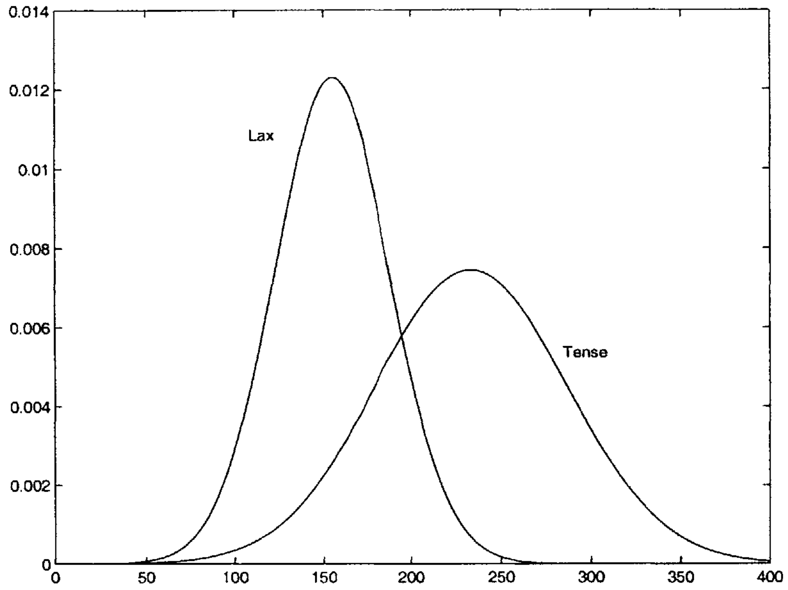

Figure 3-4: The probability of durations of tense and lax vowels after compensation. The overlap between the two distribution was reduced.

/ʌ/ followed by /r/. So we do not have to check for such a case. However, /ɝ/ may follow /r/ in some special cases (i.e. nearer, error, etc), but such /ɝ/ is mostly at the end of a word (exception: peroration), and none of them is followed by a consonant. As this thesis is dealing with the case when a vowel is surrounded on both sides by consonants, such a case need not be considered. Therefore, it may concluded that if /r/ is adjacent to a vowel, the vowel cannot be a retroflex.

To check the retroflex vowel, the F3 frequency at the midpoint of each vowel was measured. The result is shown in Figure 3-5. F3 vs B3 is plotted for better viewing.

For the retroflex vowel, F3 is significantly lower than for other vowels. From the plot of F3 vs B3, if F3 is less than 2000Hz, it may said that the vowel is almost certainly a retroflex, and if it is more than 3000Hz, it may be concluded that the vowel cannot be a retroflex.

There are some retroflex vowels for which F3 is between 2000Hz and 3000Hz. But such vowels can discriminated in the F3 vs B3 plot, because in this plot, if a retroflex has a higher F3 frequency than usual, it tends to have higher bandwidth than non-retroflex vowels. This may be because the tongue tip was not raised fully, and so

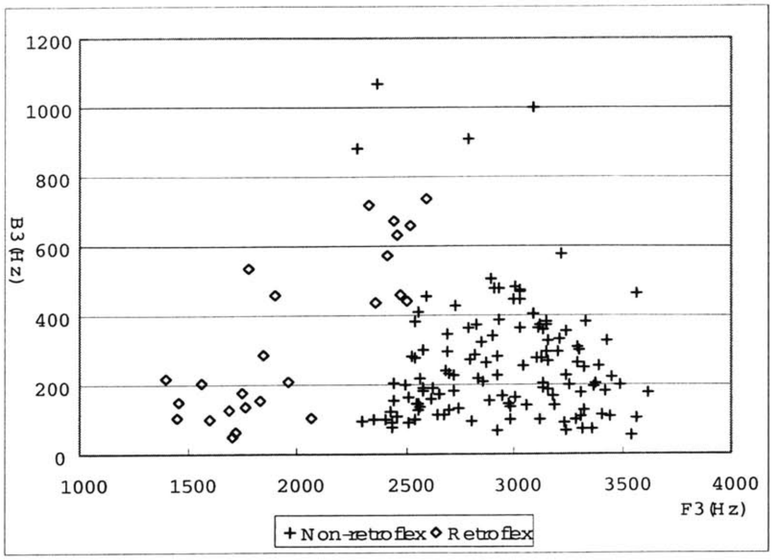

Figure 3-5: F3 vs B3 Plot of retroflex and non-retroflex vowels. They are well separated from each other without any compensation.

$\mu_{\text{retro}} = (1984, 336)$	$\mu_{\text{non}} = (2940, 259)$
$\Sigma_{\text{retro}} = \begin{pmatrix} 147971 & 71777 \\ 71777 & 52359 \end{pmatrix}$	$\Sigma_{\text{non}} = \begin{pmatrix} 109286 & -2378 \\ -2378 & 30627 \end{pmatrix}$

Table 3.7: Means and covarianaces of F3 and B3 for retroflex and non-retroflex vowels

the pole and zero that were introduced by raising of tongue tip became close to each other.

Therefore, retroflex and non-retroflex vowels can be separated rather clearly by representing each of them as a Gaussian probability distribution function. The means and the covariance matrices are shown in Table 3.7. And the contours of the probability distributions are overlayed on the F3 vs B3 plot in Figure 3-6. In this way, the retroflex vowel can be classified rather clearly.

3.5 High/Low Vowel

A high vowel is a vowel that is pronounced with a high tongue position, and low vowel is a vowel that is pronounced with a low tongue position. The vocal tract shape can be approximated very roughly as a concatenation of two uniform tubes, as in Figure

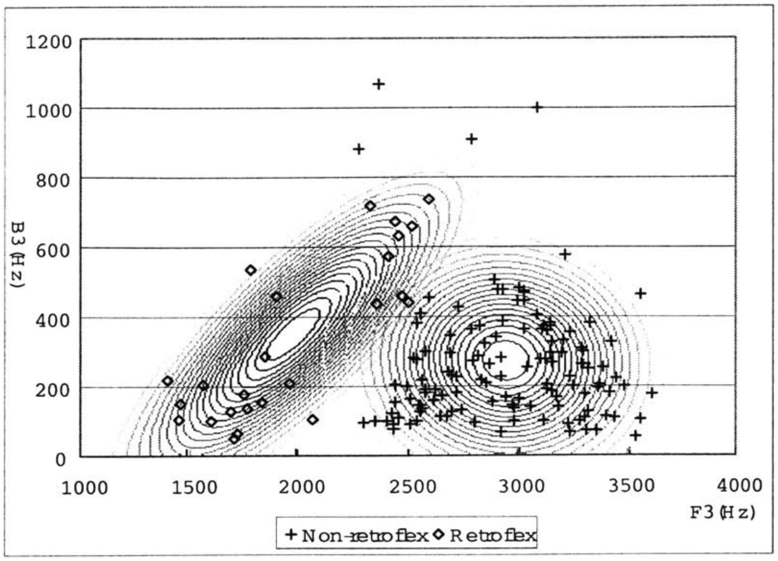

Figure 3-6: F3 vs B3 plot of retroflex and non-retroflex vowels with probability distributions overlayed

3-7. For the low vowel case, the natural frequency of each component tube affects the formant frequency of the vowel. But for the high vowel, an additional frequency, called the Helmholtz frequency appears. This frequency is determined as

$$F_{\text{Helmholtz}} = \frac{c}{2\pi}\sqrt{\frac{A_2}{A_1\ell_1\ell_2}}$$

The value of the Helmholtz frequency is usually very low, so this determines the first formant frequency of a high vowel. Therefore, if a vowel is a high vowel, F1 value goes down significantly.

The F1 frequency of each vowel has been measured. But not all of eleven vowels are measured. It would be better not to include /ɝ/ in our measurement. That is

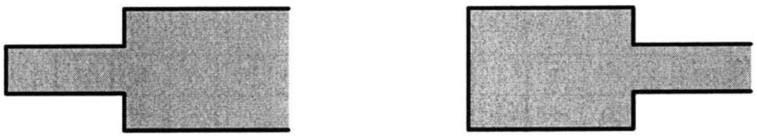

Figure 3-7: Concatenated tube representation of a low vowel(left) and high vowel(Right)

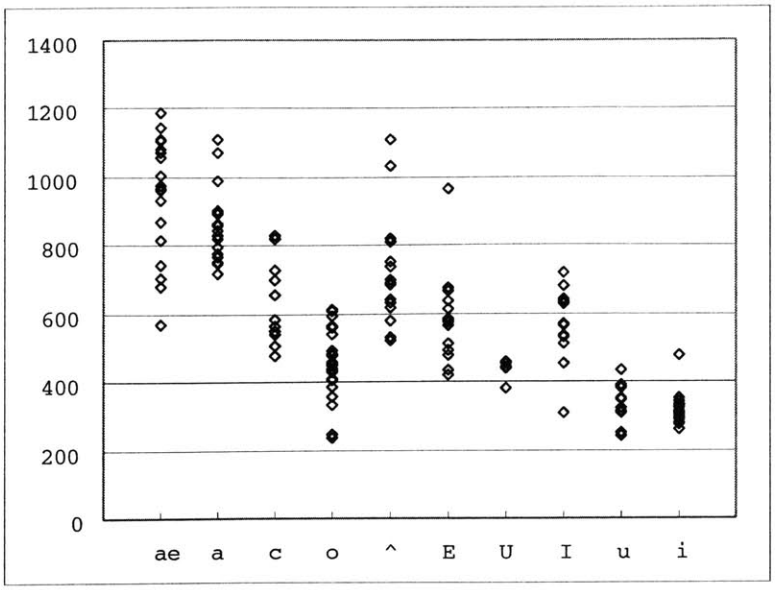

Figure 3-8: F1 values of English vowels

because the formant frequency for this vowel changes a lot from speaker to speaker, and we will need to use F3 and F4 frequencies to normalize the formant frequency. But if the same algorithm is applied to a retroflexed vowel, because F3 of a retroflex is significantly lower than usual, the normalization would lead to poor approximations to the formant frequencies. Thus, without /ɝ/, ten vowels are left. The high/low features of those ten vowels are tabulated in Table 3.8.

The F1 values of the vowels in this set have been measured. The formant frequencies are sampled at the middle of the vowel duration. Without any further processing, the result is plotted in Figure 3-8.

The F1 of /o/ is somewhat lower than expected, possibly because most of /o/ sounds are adjacent with the consonant /r/. Such an effect should be taken care of later. Looking at the figure, /æ/ and /ɑ/ are separated quite clearly from /i/ and

	i	ɪ	ɛ	æ	ɑ	ɔ	o	ʌ	u	ʊ
High	+	+	-	-	-	-	-	-	+	+
Low	-	-	-	+	+	-	-	-	-	-

Table 3.8: High/low features of English vowels

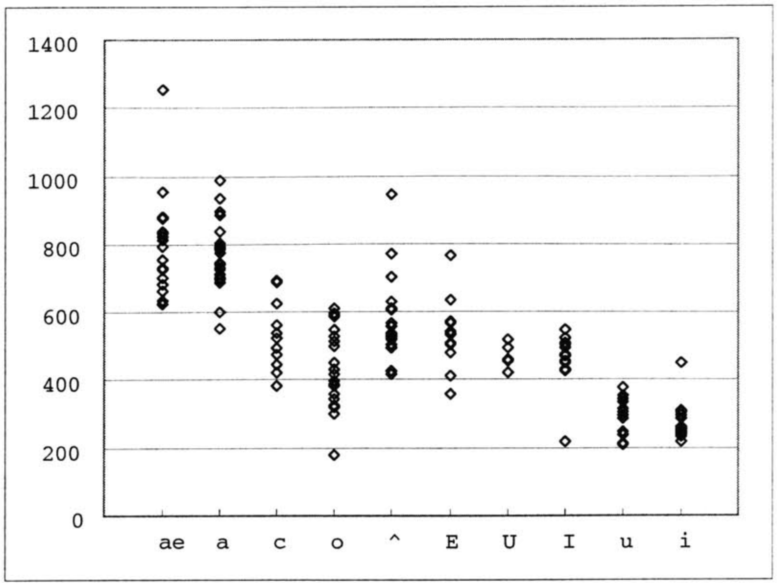

Figure 3-9: F1/F3 values of English vowels. The variances of the low vowels (/æ/, /ɔ/) are reduced.

/u/, but it would give a better result if the variance could be reduced more.

(a) Normalizing the formants.

As was stated before, the frequency of F1 depends on the configuration of the mouth, and it varies a lot depending on the speaker. So the characteristics of the speakers should be derived from the speech signal and it should be applied to normalize the F1 frequency. When we model the vocal tract by two tubes, the high-low characteristic depends mostly on the ratio of the cross-sectional area of the tubes. Therefore, some compensation for the length of the tubes needs to be made.

Generally, the third and the fourth formant frequencies do not vary a lot with time. Therefore, the third and the fourth formant frequencies may be used to compensate for the tube length. The normalization was done by using the three methods listed below.

1. $F_i = F_i/F_3$
2. $F_i = F_i/F_4$
3. $F_i = F_i/\sqrt{F_3 F_4}$

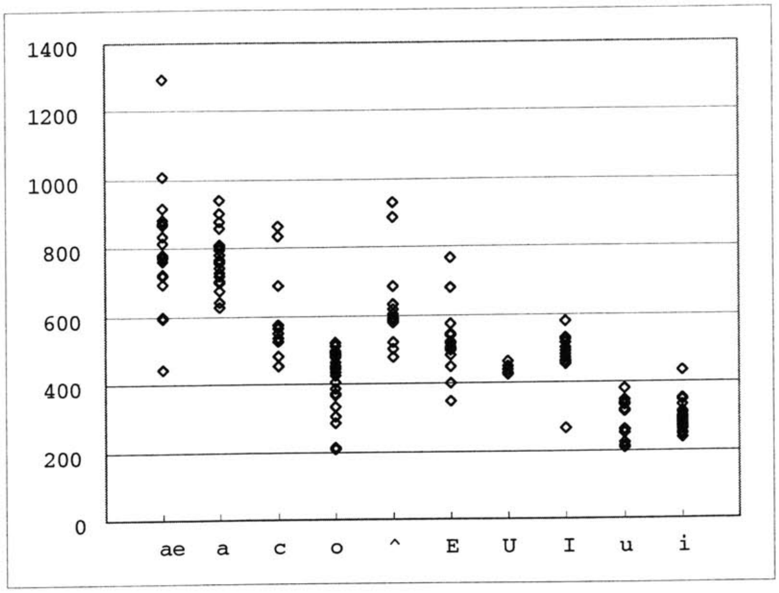

Figure 3-10: F1/F4 values of English vowels. The variances of the high vowels (ɪ, ʊ) are reduced.

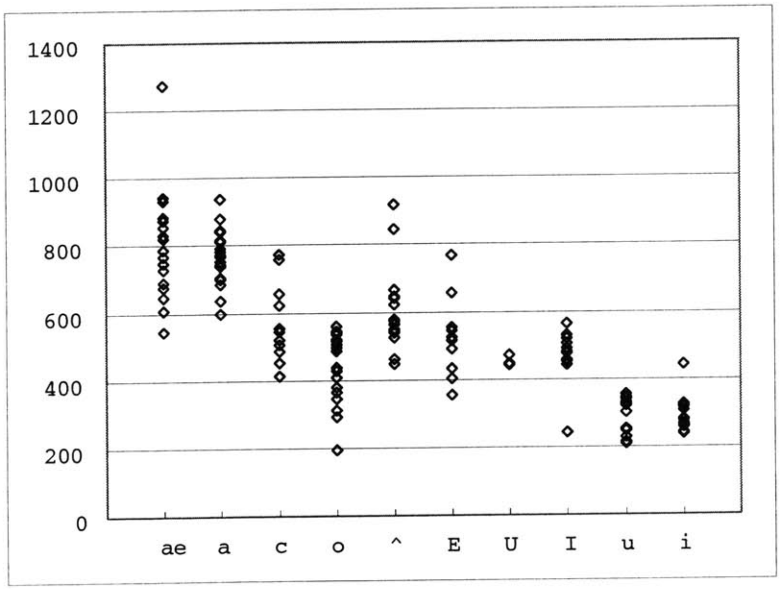

Figure 3-11: $F_1/\sqrt{F_3F_4}$ values of English vowels. This shows the best performance among the three proposed methods.

	æ	ɑ	o	ɔ	ʌ	ɛ	ʊ	ɪ	u	i
F_1	950	841	632	453	712	585	436	571	331	314
F_1/F_3	805	774	531	443	577	528	470	459	293	263
F_1/F_4	802	769	615	416	628	524	441	477	297	297
$F_1/\sqrt{F_3F_4}$	802	770	571	428	601	525	455	468	295	279

Table 3.9: Comparison of the means of the compensated F1 values. By compensating the formants, the means were reduced a little.

	æ	ɑ	o	ɔ	ʌ	ɛ	ʊ	ɪ	u	i	sum
F_1	175	103	123	99	155	135	31	107	58	43	1032
F_1/F_3	142	101	103	122	125	101	38	81	54	47	918
F_1/F_4	176	79	136	88	115	104	15	72	56	43	888
$F_1/\sqrt{F_3F_4}$	157	77	118	101	116	101	14	76	52	44	861

Table 3.10: Comparison of standard deviations of the compensated F1 values. The overall deviation is the least for $F_1/\sqrt{F_3F_4}$.

The result of each method is plotted in Figure 3-9, Figure 3-10, and Figure 3-11 respectively. All of the plots gave much better results for high vowels. The vowels /i/ and /u/ were more concentrated than before, especially the variance of /ɪ/ became significantly less. And for the low vowels, dividing by $\sqrt{F3F4}$ gave the best result.

The means and the standard deviations of the ten vowels are tabulated in Table 3.9 and Table 3.10 respectively. Table 3.10 shows that the overall standard deviation is the smallest for $F_1/\sqrt{F_3F_4}$.

(b) Compensation for Consonants

The glides and liquids affect the vowel a lot more than other consonants. Therefore the position where F1 is sampled needs to be moved farther from the consonant, to get a better estimation of F1 formant frequency value. This made the variance larger than before, but the mean could be adjusted at a more desired position.

The formant frequency depends on the speaker, but it also varies depending on the adjacent consonants. For example, if /y/ or /w/ is adjacent to a vowel, the F1 value of the vowel would be lower than usual. And when /r/ is near, F1 also tends to be lower. For stop consonants, the F1 frequency changed a little, but the influence of the consonants does not spread throughout the vowel. Consequently, the middle of

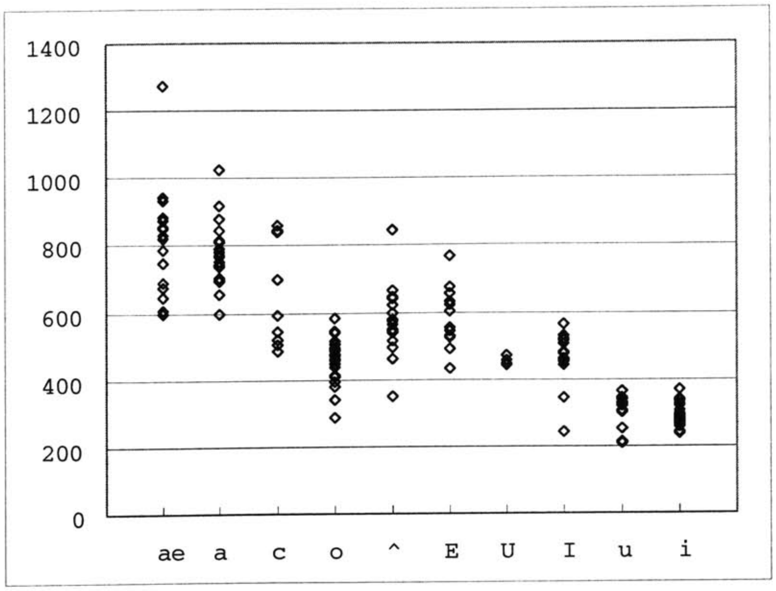

Figure 3-12: $\frac{F1}{\sqrt{F3F4}}$ values of English vowels with consonant effects compensated

the vowel is not influenced significantly. Therefore, the adjustment of the frequency was made solely for liquids and glides. In fact, the adjustment was useless for glides. F1 value was increased by 18Hz when /l/ was adjacent, and was increased by 77Hz when /r/ was adjacent. After compensating for these values, the result is as in Figure 3-12

In this figure, /i/ and /u/ are well separated from the other values. So they are classified as high vowels. Also, /æ/ and /ɑ/ are separated from the other values. They are classified as low vowels. The Gaussian distributions for High/Mid/Low vowels are given in Figure 3-13.

3.6 Front/Back Vowel

Now the case of front and back vowel can be inverstigated. The F2 formant frequency depends a lot on this feature. When the tongue position is backed, F2 is low, and when the tongue is to the front, F2 is high. The front/back feature of each monophthong is given in Table 3.11.

Now the F1 value of each vowel is measured. Without any processing, F2 for each

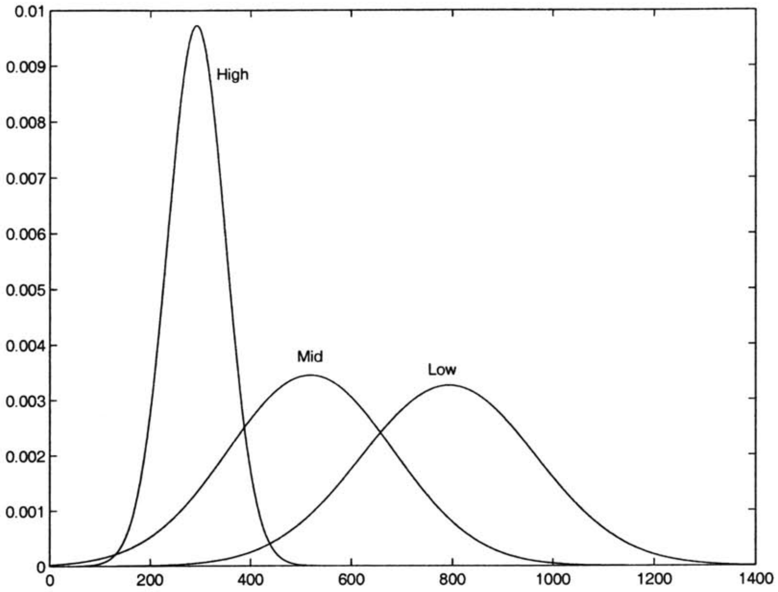

Figure 3-13: Gaussian distributions of high/mid/low vowels based on the value of $\frac{F1}{\sqrt{F3F4}}$ with consonant effects compensated

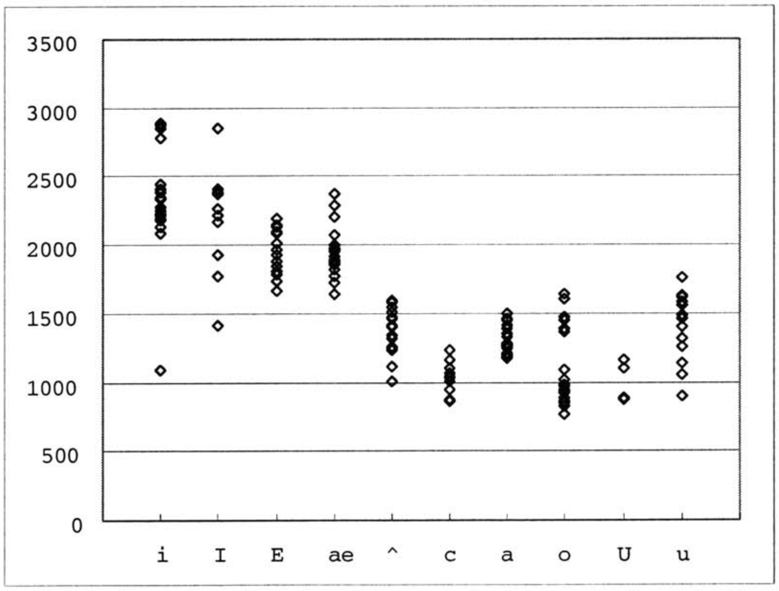

Figure 3-14: F2 values of English vowels without compensation. There are clear separation between front and back vowels

vowel is plotted in Figure 3-14. There are several very low F2 values for /i/ and /ɪ/, because of the effect of adjacent /l/. These should be taken into account later. Looking at the plot, we observe that the front/back feature is quite discriminative based on F2, without any pre-processing. But we shall try to normalize this, too.

(a) Normalizing the formants

When a vocal tract is modeled by a concatenation of two tubes, the high/low feature depends on the cross-sectional area function of the tube, so the length of the tubes were normalized as in the previous section. But for this case, the front/back feature depends on the length of each tube, so the lengths of the component tubes should not be normalized. Instead, the total length of the tube should be normalized, while maintaining the proportion.

Therefore, a simple modifying equation was set up.

$$F_{new,i} = kF_{old,i} \text{ with constant } k$$

and the constant k was determined so that F3 and F4 are close to 2500Hz and 3500Hz. The frequency 2500Hz and 3500Hz came from the analysis of uniform tube model. The value of k that minimizes the following is determined:

$$(F_{new,3} - 2500)^2 + (F_{new,4} - 3500)^2$$

The normalized second formant frequency is given by

$$F_{new,2} = F_2 \frac{2500F_3 + 3500F_4}{F_3^2 + F_4^2}$$

.

	i	ɪ	ɛ	æ	ɑ	ɔ	o	ʌ	u	ʊ
Front/Back	F	F	F	F	B	B	B	B	B	B

Table 3.11: Front/back features of English vowels

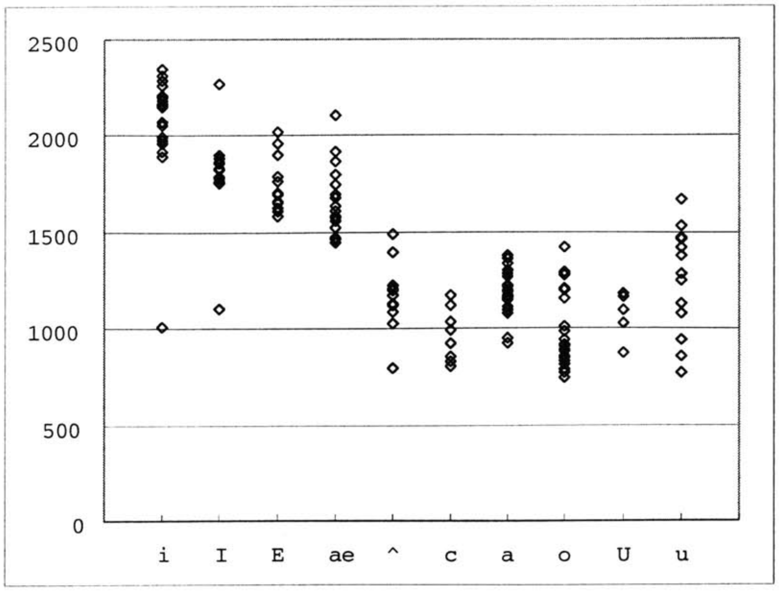

Figure 3-15: Normalized F2 values of English vowels

	æ	ɑ	o	ɔ	ʌ	ɛ	ʊ	ɪ	u	i
F_2	2317	2191	1949	1960	1369	1032	1301	1098	1039	1412
$F_{new,2}$	2073	1807	1752	1648	1172	947	1188	1004	1068	1265

Table 3.12: Comparison of the means of compensated F2 values

The normalized F2 values are plotted in Figure 3-15. It can be seen that it worked very well for the front vowels. The actual means and the standard deviations are tabulated in Table 3.12 and Table 3.13.

(b) Compensation for Consonants

For the front/back case, too, the effects of the consonants need to be compensated. The compensation was made for the liquids and glides. But again this time, the glides did not make significant difference. So 10Hz was added for initial /r/, 218Hz was added for final /r/. Also 71Hz was added for Initial /l/, and 262Hz was added

	æ	ɑ	o	ɔ	ʌ	ɛ	ʊ	ɪ	u	i	sum
F_2	370	348	168	184	159	112	108	283	147	248	2131
$F_{new,2}$	268	247	132	173	141	122	114	215	125	272	1814

Table 3.13: Comparison of standard deviations of compensated F2 values

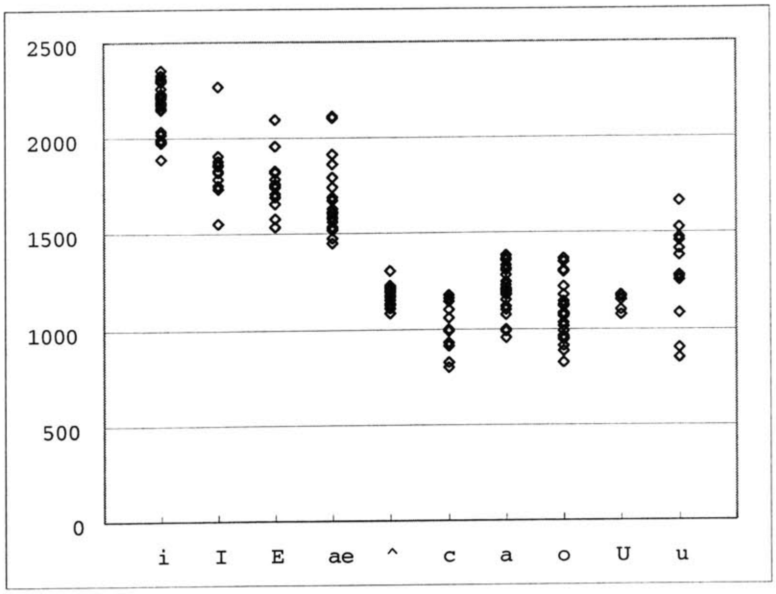

Figure 3-16: Normalized F2 values of English vowels with consonant effect compensated

	æ	ɑ	o	ɔ	ʌ	ɛ	ʊ	ɪ	u	i	sum
Mean	2165	1868	1760	1684	1172	1007	1197	1066	1134	1291	
SD	129	199	142	195	53	131	124	149	44	238	1407

Table 3.14: Mean and standard deviation of consonant compensated F2 values. The overall variation was reduced significantly after compensation.

for Final /l/. The plot of the compensated value is given in Figure 3-16.

After compensating for the consonants, most of the values that were far away from the mean moved closer to the mean value of each vowel. Also the consonant compensation gives a smaller variance to the back vowels. The means and standard deviations are given in Table 3.14

Using these corrections, we can determine the front/back feature as in Figure 3-17.

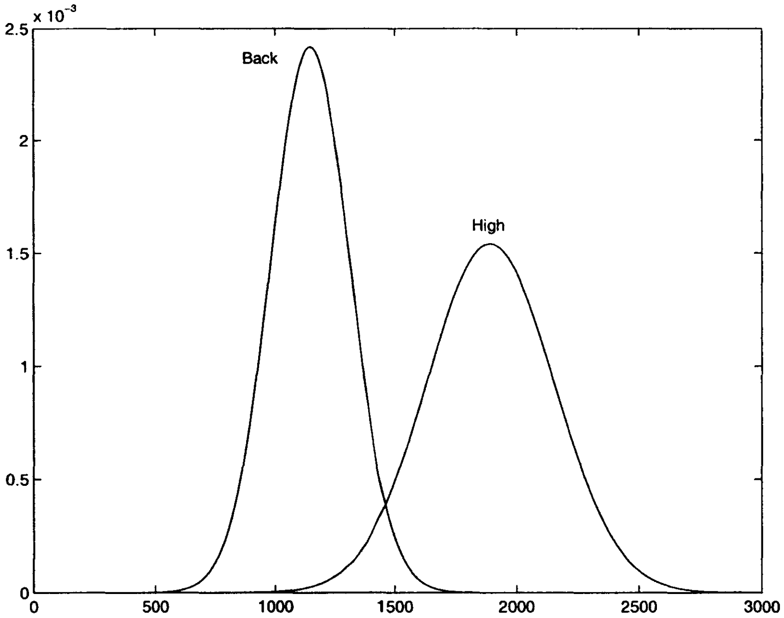

Figure 3-17: Normalized F2 values of English vowels with consonant effects compensated

Chapter 4

Recognition

Based on the analyzed data in Chapter 3, each of thc vowels will be recognized in a different manner.

4.1 Recognition of Each Features

In Chapter 3, four different features were evaluated — Tense/Lax, High/Low, Front/Back, and Retroflex. In this section, the performance of the recognition of each feature is evaluated. The database to be recognized consists of 1000 words that are randomly selected from the database made from Chapter 3. The same labels are used. The formant track is not hand corrected for this case.

4.1.1 Recognition of Tense/Lax Feature

The tense/lax feature is determined in part by the duration of the vowels. The probability of being a tense or lax vowel depending on the duration was calculated from the probability distributions $P(d|\text{Tense})$ and $P(d|\text{Lax})$ evaluated in Chapter 3, using the following relationship.

$$P(\text{Tense}|d) = \frac{P(d|\text{Tense})P(\text{Tense})}{P(d)}$$

$$P(\text{Lax}|d) = \frac{P(d|\text{Lax})P(\text{Lax})}{P(d)}$$

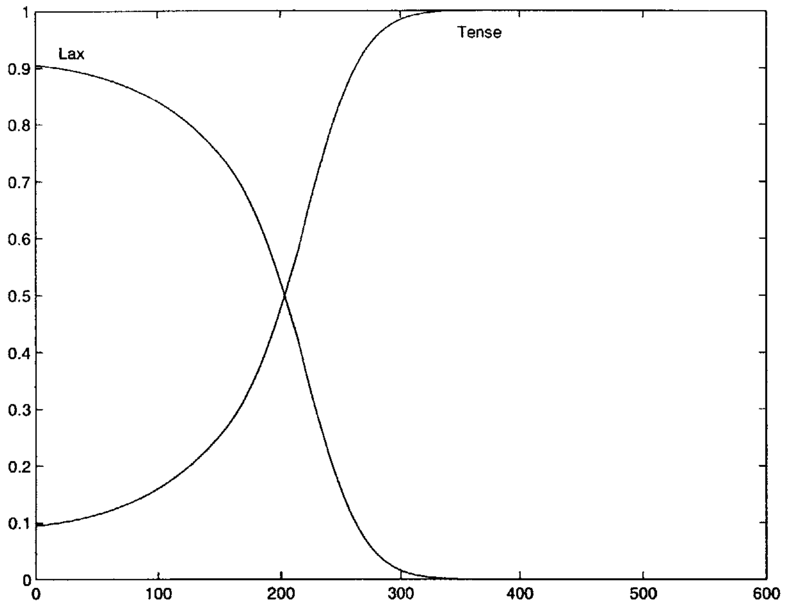

Figure 4-1: The probability of tense/lax features according to the vowel duration

And, as we should determine $P(\text{Tense}|d)$ and $P(\text{Lax}|d)$, so that $P(\text{Tense}|d)+P(\text{Lax}|d) = 1$, the following relationship can de derived.

$$
\begin{aligned}
P(\text{Tense}|d) &= \frac{P(d|\text{Tense})P(\text{Tense})}{P(d|\text{Lax})P(\text{Lax}) + P(d|\text{Tense})P(\text{Tense})} \\
P(\text{Lax}|d) &= \frac{P(d|\text{Lax})P(\text{Lax})}{P(d|\text{Lax})P(\text{Lax}) + P(d|\text{Tense})P(\text{Tense})}
\end{aligned}
$$

The evaluated probability distributions are shown in Figure 4-1. This figure shows that if a vowel duration is longer than 300ms, it can be said to be a tense vowel with confidence, but if the vowel duration gets short, there is still quite a large probability that it may not be a lax vowel.

The confusion matrix of this feature is shown in Table 4.1. If a vowel is determined to be a tense vowel, this can be correct with a probability of more than 96.9%. But when a vowel is determined to be a lax vowel, this can be correct with a probability of only 59.5%. The overall error rate is 24.6%. Therefore, we can determine the tense/lax feature with more than 75% accuracy.

	Tense	Lax
Tense	411	233
Lax	13	343

Table 4.1: Confusion matrix of tense/lax features of 1000 English vowels. The labels on the left are the actual features, and the labels on top are the recognized features.

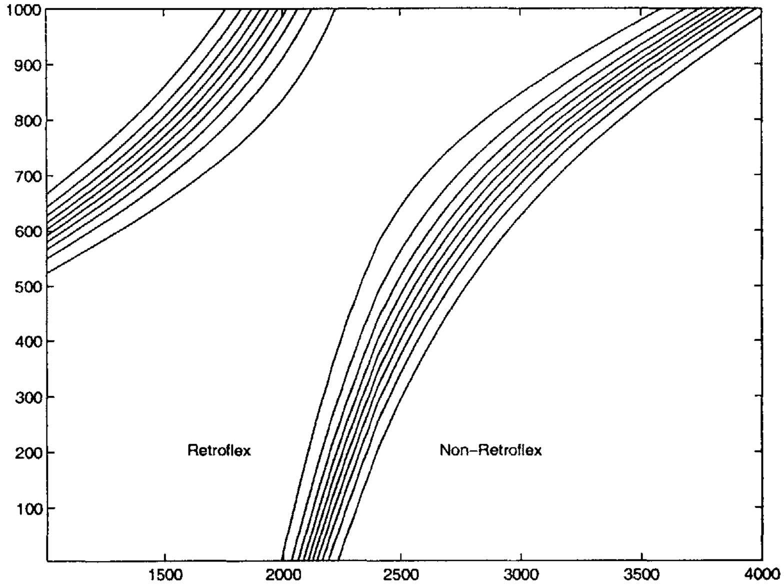

Figure 4-2: The contour plot of the probability of retroflex features according to F3 and B3. The x-axis represents the F3 values, and y-axis represents B3 values. Each contour represents a different probability from 0.1 to 0.9.

4.1.2 Recognition of Retroflex Feature

The retroflex feature is determined by the F3 value. The probability of being a retroflex is plotted as contours in Figure 4-2. The probability is calculated using the following relations.

$$P(\text{Retro}|d) = \frac{P(d|\text{Retro})P(\text{Retro})}{P(d|\text{Non})P(\text{Non}) + P(d|\text{Retro})P(\text{Retro})}$$
$$P(\text{Non}|d) = \frac{P(d|\text{Non})P(\text{Non})}{P(d|\text{Non})P(\text{Non}) + P(d|\text{Retro})P(\text{Retro})}$$

The boundary between the retroflex and non-retroflex is very abrupt, so that retroflex vowels can be determined with high confidence.

	Retroflex	Non-retroflex
Retroflex	80	13
Nonretroflex	38	681

Table 4.2: Confusion matrix of retroflex features of 812 English vowels. The labels on the left are the actual features, and the labels on top are the recognized features.

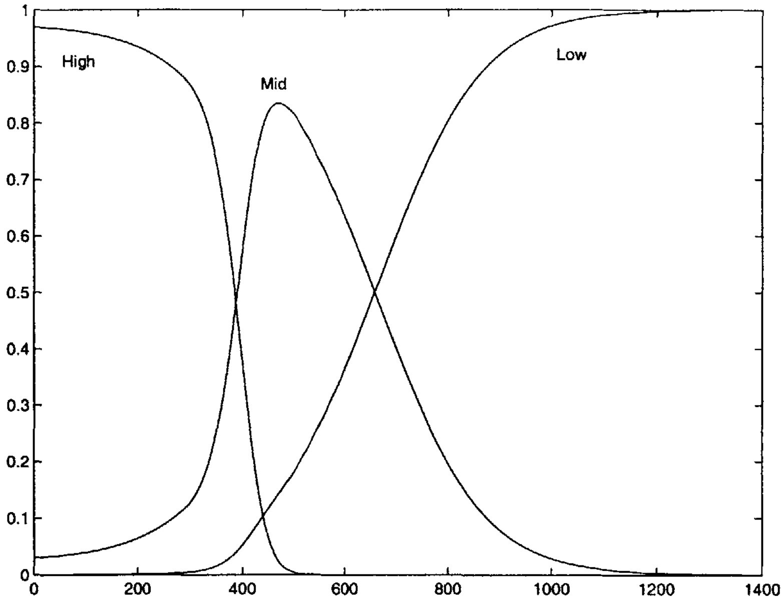

Figure 4-3: The probability of high/low features with respect to normalized F1

The confusion matrix of this feature is given in Table 4.2. The overall accuracy is more than 93%. The recognition of the retroflex was quite successful. When a vowel is determined to be a non-retroflex, it is correct with more than 98% accuracy. And if a vowel is determined to be a retroflex, it is correct with just about 67.7% accuracy. But most of the wrongly detected retroflex were at the boundary of a retroflex and a non-retroflex, and they also have large probability of being non-retroflex.

4.1.3 Recognition of High/Low Feature

The high/low feature is determined by F1 value, and it is classified into three different classes — High, Mid, Low. High includes /i/ and /u/, Low includes /æ/ and /ɑ/. The probability of each class is plotted in Figure 4-3.

	High	Mid	Low
High	193	5	1
Mid	50	373	58
Low	0	35	192

Table 4.3: Confusion matrix of high/low features of 907 English vowels. The labels on the left are the actual features, and the labels on top are the recognized features.

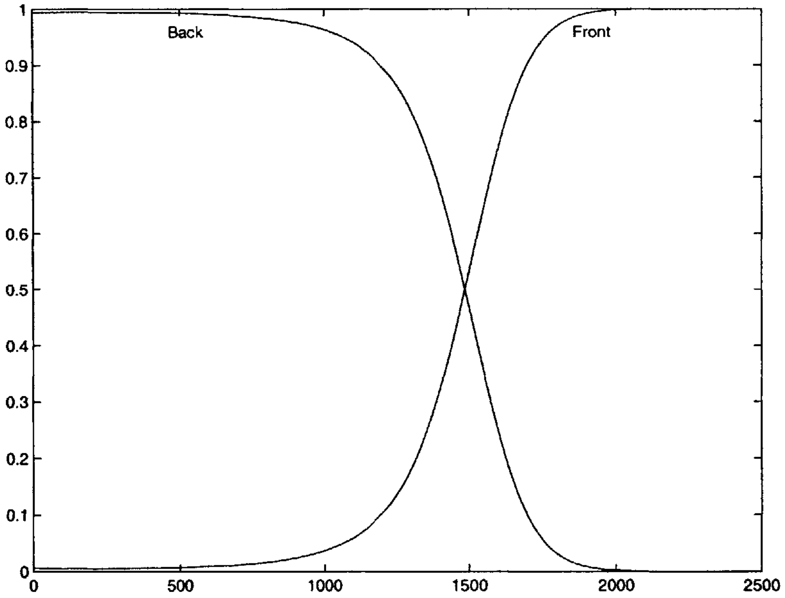

Figure 4-4: The probability of front/back features with respect to normalized F2

The confusion matrix of this feature is given in Table 4.3. With a few exceptions, a high vowel is not recognized as a low vowel, and a low vowel is not recognized as a high vowel. Most of the errors are made between Low/Mid or Mid/High. A high vowel is recognized as a high vowel with a great accuracy. But the recognition rate of the mid and low vowels are not that good. The overall recognition rate is 83.6%.

4.1.4 Recognition of Front/Back Feature

The front/back feature is determined by using F2 value, and in Chapter 3 it was discriminated with a quite good accuracy. The probability of each class is plotted in Figure 4-4.

The confusion matrix for this feature is given in Table 4.4. It could be recognized

	Front	Back
Front	360	13
Back	17	517

Table 4.4: Confusion matrix of front/back features of 907 English vowels. The labels on the left are the actual features, and the labels on top are the recognized features.

Features	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	o	ʊ	u	ɝ
Tense/Lax	T	L	L	T	T	T	L	T	L	T	T
Retroflex											R
High/Mid/Low	H	M	M	L	L	M	M	M	M	H	-
Front/Back	F	F	F	F	B	B	B	B	B	B	-

Table 4.5: Features of English vowels. ɪ and ʊ are usually classified as high vowels, but in this thesis, it was classified as mid vowels to get a better classification results.

with a great accuracy. The overall accuracy is about 97%.

4.2 Order of Recognition

4.2.1 Problem of Feature Based Recognition

Fron the probability of identifying each feature for a vowel, it is possible to estimate the probability of being each particular vowel. Look at Table 4.5. By knowing the features of each vowel, we can determine the vowel. For example, if a vowel has Tense, High, and Back features, it is likely to be the vowel /u/. But there are several problems with this. If a vowel was determined to have Tense, Mid, and Back features, it can be either /ɔ/ or /o/. We need further processing to determine which is more correct. And if a vowel was determined to have Tense, Mid, and Front features, there is no vowel corresponding to this, at lease in this study. We need to consider these in determining the vowel to avoid such problems.

Let us summarize the problems. First, there are multiple vowels which have same sets of features. They are tabulated in Table 4.6. We can see that this problem arises because the number of vowels that are contained in Mid feature is too large. Thus, we may be able to solve this problem by dividing the Mid class into smaller classes.

Look at Figure 3-12 of Chaper 3. The vowels in Mid class can be classified into two groups — /ɪ/, /ʊ/, o, vs. /ɔ/, /ʌ/, /ɛ/. Therefore, by doing so, this incomplete classification problem can be solved.

Features	Vowel
Lax/Mid/Front	/ɪ/, /ɛ/
Lax/Mid/Back	/ʊ/, /ʌ/
Tense/Mid/Back	/ɔ/, /o/

Table 4.6: Vowels with the same sets of features. The vowels with mid feature have this problem.

Also, many sets of features do not have a corresponding vowel. The sets of features are — Tense/Mid/Front, Lax/High/Front, Lax/High/Back, Lax/Low/Front, Lax/Low/Back, Retroflex/Lax. We can see that for Lax vowels, they always have Mid feature. And retoflex vowels mostly have a longer duration, so they should be Tense. In such cases, we may simply discard the case, and look for the next highest probability.

4.2.2 Recognition Hierarchy

We first determine what order of recognition to select in order to give the highest possible performance. First of all, the retoflex feature should be determined. This is because we should not use a retroflex in determining the High/Low feature, or the Front/Back feature. Thus we get the following probabilities.

$$P(\text{ɝ}),\ P(\neg\text{ɝ})$$

For the vowels that are recognized as a retroflex, we check the Tense/Lax feature to see if it's really a retroflex. And for the vowels that are recognized as a non-retroflex, we go on to the second stage. At this stage, we may determine the feature with the highest accuracy, which is Front/Back feature. By knowing if it is a front vowel, or a back vowel, we can get the following two probabilities.

$$\begin{aligned} P(\text{Front}) &= P(\text{Front}|\neg\text{ɝ})P(\neg\text{ɝ}) \\ P(\text{Back}) &= P(\text{Back}|\neg\text{ɝ})P(\neg\text{ɝ}) \end{aligned}$$

And for the next stage, the high/low feature is determined, mainly because the tense/lax feature has lower recognition rate than the high/low feature. But the probability distribution of high/low features in the previous section should be changed so that the feature is classified into four different classes. — High, Low, MidHigh, and MidLow.

The probability distrubutions of these classes are shown in Figure 4-5. In fact, it would give a better performance if different probability distributions are used for Front vowel and Back vowel cases. But because they do not differ from each other so much, and because having different distributions for every different case would take too much computation, it was assumed that

$$\begin{aligned} &P(\text{High}|\text{Front}) \simeq P(\text{High}|\text{Back}) \simeq P(\text{High}) \\ &P(\text{MidHigh}|\text{Front}) \simeq P(\text{MidHigh}|\text{Back}) \simeq P(\text{MidHigh}) \\ &P(\text{MidLow}|\text{Front}) \simeq P(\text{MidLow}|\text{Back}) \simeq P(\text{MidLow}) \\ &P(\text{Low}|\text{Front}) \simeq P(\text{Low}|\text{Back}) \simeq P(\text{Low}) \end{aligned}$$

And then the tense/lax feature should be measured. But in fact, this does not have a significant importance, because if the high/low and front/back features are determined, the tense/lax feature is almost determined. This stage is, in a sense, a stage to verify if the other features were recognized correctly. The only case tense/lax feature has an importance is MidHigh/Back and MidLow/Back cases.

But as we did not analyze the case when the vowel is adjacent to a liquid or glide, for those cases this stage cannot be evaluated. Therefore, the tense/lax feature of the case was calculated for liquids and glides. The probability distribution is plotted in Figure 4-6. As you see, this does not show much sidcrimination.

Summing all of those cases, the recognition hierachy is given in Figure 4-7

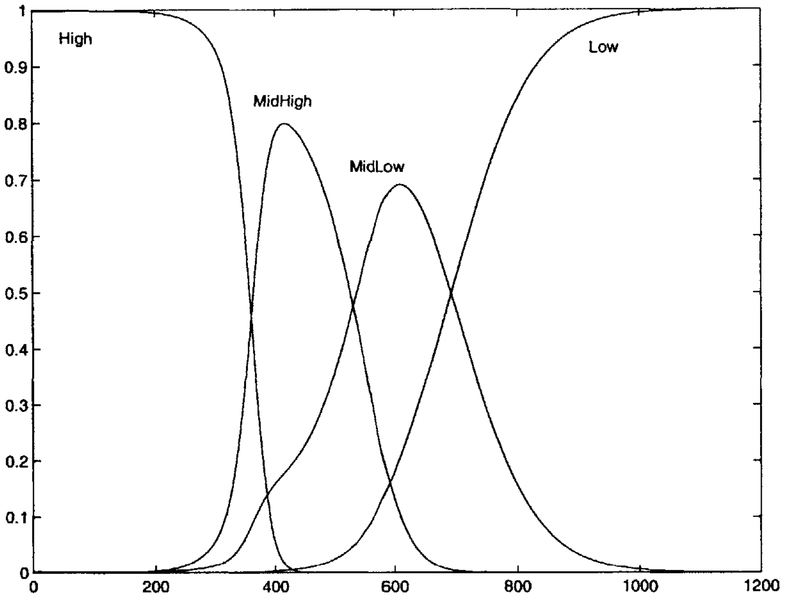

Figure 4-5: The probability of high/low features segmented into four classes — High, Midhigh, Midlow, Low

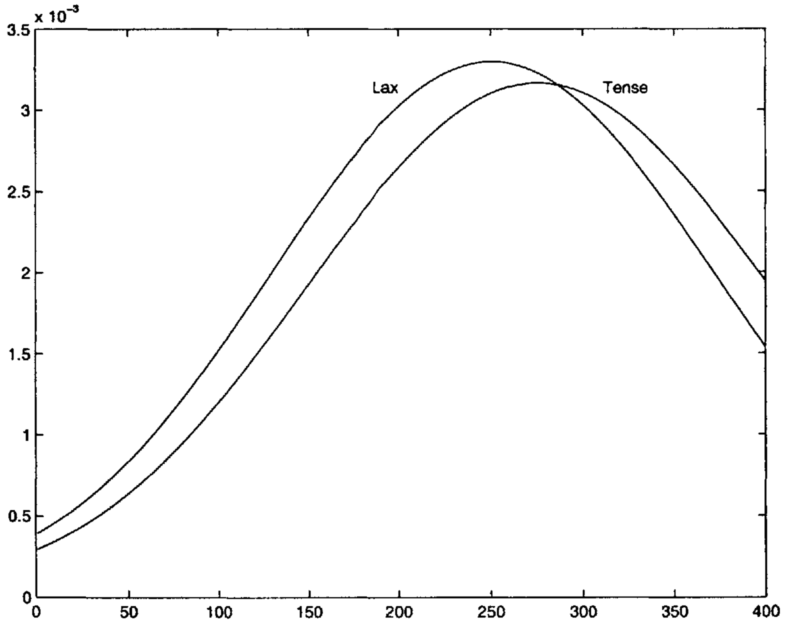

Figure 4-6: The probability of tense/lax features given that one of the consonants is a liquid or a glide. This does not show a large discrimination.

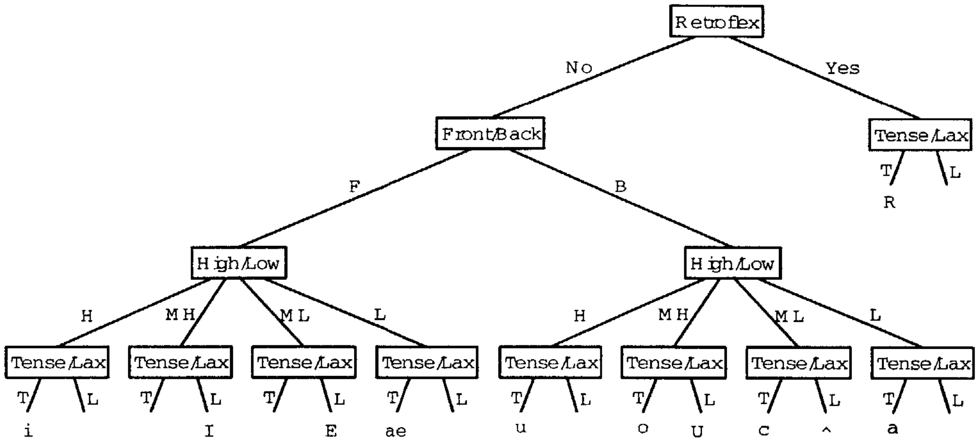

Figure 4-7: The hierarchy tree of sequential recognition. The node with no vowel will be discarded.

4.3 Recognition of Vowel without Consonant Compensated Probability

4.3.1 Recognition without Limited Candidates

First, the vowels were recognized without compensating for any special consonant effects. 1000 vowels were recognized by the method. The confusion matrix of this recognition system is given in Table 4.7. The overall recognition rate is 54.7%.

4.3.2 Recognition with Limited Candidates

In the previous section, the set of possible candidates was not considered. Therefore, if a sound was of the form /b-(vowel)-g/, the previous method may give a result /bʊg/, which cannot be a word. Therefore, if a recognized vowel did not make sense, it may be discarded, and the vowel with the next highest probability can be selected as the recognition result.

By this method, we get the confusion matrix in Table 4.8, and the recognition rate is 74.1%. This is a lot better than recognizing without limiting the consonant candidates.

	i	ɪ	ɛ	æ	u	o	ʊ	ɔ	ʌ	ɑ	ɝ
i	104	12	5	0	1	0	0	0	0	0	5
ɪ	12	74	46	0	0	0	0	5	9	0	2
ɛ	0	73	49	11	0	0	0	0	0	0	8
æ	0	7	10	82	0	0	0	0	0	0	19
u	0	0	0	0	40	0	0	2	14	0	16
o	0	0	2	0	16	31	23	4	2	0	0
ʊ	0	0	0	0	4	2	5	0	0	0	6
ɔ	0	0	0	0	0	14	9	6	7	10	1
ʌ	0	0	0	0	0	1	4	2	32	5	6
ɑ	0	0	0	0	0	0	0	5	16	49	39
ɝ	11	7	0	0	0	0	0	0	0	0	75

Table 4.7: Confusion matrix of recognition without consonant compensated probability distribution. The labels on the left are the actual vowels, and the labels on top are the recognized vowels.

	i	ɪ	ɛ	æ	u	o	ʊ	ɔ	ʌ	ɑ	ɝ
i	115	7	3	0	0	0	0	0	0	0	2
ɪ	4	87	45	0	0	0	1	2	9	0	0
ɛ	0	26	102	11	0	0	0	0	0	0	2
æ	0	6	7	100	0	0	0	0	0	0	5
u	0	0	0	0	61	0	0	0	4	0	7
o	0	0	1	1	7	47	19	2	1	0	0
ʊ	0	0	0	0	1	2	9	0	0	0	5
ɔ	0	0	0	0	0	11	2	28	5	1	0
ʌ	0	0	0	0	0	2	1	1	42	2	2
ɑ	0	0	0	0	0	0	0	3	14	72	20
ɝ	5	8	2	0	0	0	0	0	0	0	78

Table 4.8: Confusion matrix of vowel recognition with uncompensated probability and limited candidates. The labels on the left are the actual vowels, and the labels on top are the recognized vowels

	i	ɪ	ɛ	æ	u	o	ʊ	ɔ	ʌ	ɑ	ɝ
i	109	9	0	0	3	0	1	0	0	0	5
ɪ	25	101	17	0	0	0	1	0	1	0	3
ɛ	3	29	91	10	0	0	2	0	0	1	5
æ	0	0	11	85	0	0	0	0	0	7	15
u	8	2	0	0	46	5	0	0	0	0	11
o	0	0	0	0	6	41	23	7	1	0	0
ʊ	0	0	0	0	0	2	11	0	0	0	4
ɔ	0	0	0	0	0	11	8	15	7	5	1
ʌ	0	0	0	0	0	6	1	3	36	2	2
ɑ	0	0	0	0	0	0	0	10	5	75	19
ɝ	4	7	0	0	0	0	0	0	0	0	82

Table 4.9: Confusion matrix of vowel recognition with consonant compensated probability distribution. The labels on the left are the actual vowels, and the labels on top are the recognized vowels.

4.4 Recognition of Vowel with Consonant Compensated Probability

4.4.1 Recognition without Limited Candidates

In this section, the compensated probability will be used to recognize the vowel. First, we use the compensated probability calculated in Chapter 3, without considering limiting the candidates by consonant context. The confusion matrix of this recognition system is given in Table 4.9. The overall recognition rate is 69.6%. This shows an improvement about 15% compared to using uncompensated probability.

4.4.2 Recognition with Limited Candidates

Now, the previous method was adjusted to account for limitation of the number of candidates by the consonant context. Using this method, we get the confusion matrix in Table 4.10, and the recognition rate is 87.4%. This shows the best result compared with the other three methods.

	i	ɪ	ɛ	æ	u	o	ʊ	ɔ	ʌ	ɑ	ɝ
i	118	7	0	0	1	0	0	0	0	0	1
ɪ	12	129	6	0	0	0	1	0	0	0	0
ɛ	1	14	119	5	0	0	0	0	0	1	1
æ	0	0	4	107	0	0	0	0	0	3	4
u	2	0	0	0	62	3	0	0	0	0	5
o	0	0	0	0	2	61	13	2	0	0	0
ʊ	0	0	0	0	0	1	13	0	0	0	3
ɔ	0	0	0	0	0	4	1	39	1	2	0
ʌ	0	0	0	0	0	1	1	0	47	1	0
ɑ	0	0	0	0	0	0	0	4	3	90	12
ɝ	0	4	0	0	0	0	0	0	0	0	89

Table 4.10: Confusion matrix of vowel recognition with consonant compensated probability distribution and limited candidates

4.5 Summary

If we know the consonants adjacent to a vowel, we can either use a probability density function that compensates for the effects of the consonants, or limit the number of candidates using the consonantal context. In this section, such different methods of recognition were used to recognize a vowel. Figure 4-8 compares the recognition accuracy of different methods. As can be expected, using both methods gave the highest performance.

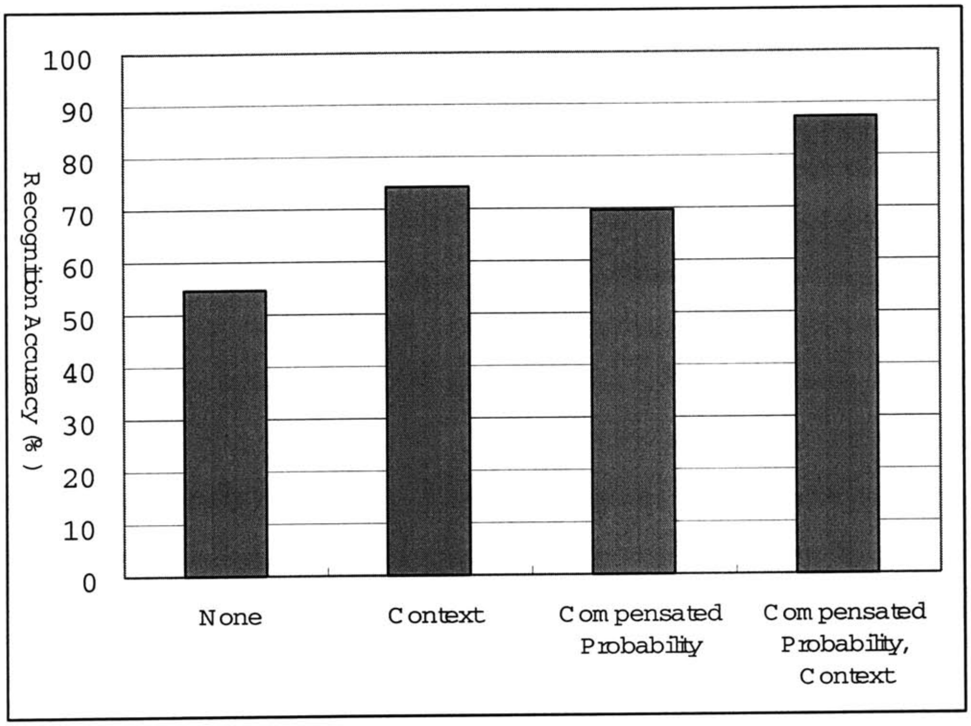

Figure 4-8: Recognition results for four different experiments.

Chapter 5

Conclusion

The top-down method is a method which uses a information of higher level to determine the lower level. In this thesis, a small fraction of the speech recognition system with top-down method was implemented. By applying a top-down method, we could have the advantage that we can use the previously recognized information to improve the performance, and that we can limit the number of candidates using the information from the higher level.

But there is still a lot of further research that needs to be done. First, the formants can be tracked better by using the consonantal context. For example, a nasal usually has a higher tilt and extra peaks in the spectrum, and the aspiration noise can interfere with formant tracking. If we can compensate for these properties, we may get a better formant track. And in this thesis, only one point was sampled within each vowel to estimate the features. But if we can consider the temporal movement of the formants, it would give more information of compensating for the consonants. Also, the order of recognition was the same for all cases, but this can be made more flexible. For example, if there are no retroflex vowels in the candidate words, we do not need to check the retroflex feature.

This thesis dealt with the interaction between the word level and the phoneme level, especially the vowels. Similar studies may be carried out to recognize consonants, or the interaction between other levels of speech can be used to improve the overall quality of the recognition system.

Appendix A

List of words used in the database

batch	/bæʧ/	wrath	/ræθ/	rung	/rʌŋ/	yon	/yɑn/
bag	/bæg/	sang	/sæŋ/	sung	/sʌŋ/	czar	/zɑr/
chap	/ʧæp/	tam	/tæm/	shut	/ʃʌt/	bog	/bɔg/
fad	/fæd/	tap	/tæp/	thus	/ðʌs/	chalk	/ʧɔk/
gad	/gæd/	wear	/wær/	thud	/θʌd/	gaud	/gɔd/
ham	/hæm/	than	/ðæn/	thumb	/θʌm/	gone	/gɔn/
has	/hæz/	there	/ðær/	botch	/bɑʧ/	moth	/mɔθ/
jab	/ʤæb/	bug	/bʌg/	chock	/ʧɑk/	naught	/nɔt/
jag	/ʤæg/	chuck	/ʧʌk/	chop	/ʧɑp/	wrong	/rɔŋ/
cad	/kæd/	duck	/dʌk/	dock	/dɑk/	wroth	/rɔθ/
calve	/kæv/	fuzz	/fʌz/	dot	/dɑt/	song	/sɔŋ/
laugh	/læf/	gun	/gʌn/	god	/gɑd/	war	/wɔr/
lash	/læʃ/	hum	/hʌm/	job	/ʤɑb/	yawl	/yɔl/
mare	/mær/	jug	/ʤʌg/	jog	/ʤɑg/	yawn	/yɔn/
math	/mæθ/	cud	/kʌd/	cod	/kɑd/	thong	/θɔŋ/
nab	/næb/	luff	/lʌf/	mar	/mɑr/	beg	/bɛg/
gnat	/næt/	lush	/lʌʃ/	knob	/nɑb/	check	/ʧɛk/
patch	/pæʧ/	nub	/nʌb/	not	/nɑt/	deck	/dɛk/
pass	/pæs/	nut	/nʌt/	shot	/ʃɑt/	debt	/dɛt/
rang	/ræŋ/	pus	/pʌs/	top	/tɑp/	fed	/fɛd/

Table A.1: Words with same consonant pairs (1/2)

fez	/fɛz/	theme	/θim/	thin	/θɪn/	furze	/fɝz/
hem	/hɛm/	zeal	/zil/	zing	/zɪŋ/	gird	/gɝd/
lev	/lɛf/	wreathe	/rið/	zip	/zɪp/	hers	/hɝz/
neb	/nɛb/	seethe	/sið/	choke	/ʧok/	curd	/kɝd/
net	/nɛt/	sheathe	/ʃið/	dote	/dot/	curve	/kɝv/
sedge	/sɛʤ/	teethe	/tið/	goad	/god/	mirth	/mɝθ/
share	/ʃɛr/	bitch	/bɪʧ/	home	/hom/	perch	/pɝʧ/
yell	/yɛl/	big	/bɪg/	hose	/hoz/	purse	/pɝs/
yen	/yɛn/	chick	/ʧɪk/	code	/kod/	surge	/sɝʤ/
then	/ðɛn/	chip	/ʧɪp/	cove	/kov/	shirt	/ʃɝt/
their	/ðɛr/	fizz	/fɪz/	loaf	/lof/	term	/tɝm/
zed	/zɛd/	gin	/gɪn/	more	/mor/	verse	/vɝs/
Zen	/zɛn/	him	/hɪm/	note	/not/	yearn	/yɝn/
beach	/biʧ/	hymn	/hɪm/	poach	/poʧ/	third	/θɝd/
cheek	/ʧik/	his	/hɪz/	shore	/ʃor/	duke	/duk/
cheap	/ʧip/	jib	/ʤɪb/	shoat	/ʃot/	food	/fud/
feed	/fid/	jig	/ʤɪg/	tome	/tom/	whom	/hum/
feaze	/fiz/	kid	/kɪd/	tope	/top/	whose	/huz/
leaf	/lif/	mere	/mɪr/	vole	/vol/	newt	/nut/
leash	/liʃ/	myth	/mɪθ/	wore	/wor/	shoot	/ʃut/
neat	/nit/	nib	/nɪb/	wove	/wov/	tomb	/tum/
peach	/piʧ/	knit	/nɪt/	those	/ðoz/	yule	/yul/
peace	/pis/	nit	/nɪt/	zone	/zon/	zoom	/zum/
wreath	/riθ/	pitch	/pɪʧ/	loathe	/loð/	soothe	/suð/
siege	/siʤ/	ring	/rɪŋ/	loge	/loʒ/	rouge	/ruʒ/
sheet	/ʃit/	sing	/sɪŋ/	birch	/bɝʧ/	good	/gʊd/
team	/tim/	shear	/ʃɪr/	berg	/bɝg/	could	/kʊd/
veal	/vil/	tip	/tɪp/	chirp	/ʧɝp/	moor	/mʊr/
weave	/wiv/	this	/ðɪs/	dirk	/dɝk/	putsch	/pʊʧ/
yean	/yin/	thing	/θɪŋ/	dirt	/dɝt/	sure	/ʃʊr/
these	/ðiz/						

Table A.2: Words with same consonant pairs (2/2)

hatch	/hæʧ/	mutt	/mʌt/	league	/lig/	rogue	/rog/
had	/hæd/	cob	/kɑb/	liege	/liʤ/	role	/rol/
half	/hæf/	cod	/kɑd/	leak	/lik/	roam	/rom/
hag	/hæg/	cog	/kɑg/	leal	/lil/	roan	/ron/
hang	/hæŋ/	cock	/kɑk/	lean	/lin/	rope	/rop/
hack	/hæk/	calm	/kɑm/	leap	/lip/	roar	/ror/
ham	/hæm/	con	/kɑn/	lease	/lis/	wrote	/rot/
hap	/hæp/	cop	/kɑp/	leash	/liʃ/	rove	/rov/
hair	/hær/	car	/kɑr/	lied	/lit/	rose	/roz/
hash	/hæʃ/	cot	/kɑt/	leave	/liv/	tube	/tub/
hat	/hæt/	watch	/wɔʧ/	lees	/liz/	tuque	/tuk/
have	/hæv/	walk	/wɔk/	nib	/nɪb/	tool	/tul/
has	/hæz/	wall	/wɔl/	niche	/nɪʧ/	tomb	/tum/
much	/mʌʧ/	war	/wɔr/	nick	/nɪk/	tune	/tun/
mud	/mʌd/	wash	/wɔʃ/	nil	/nɪl/	toot	/tut/
muff	/mʌf/	yell	/yɛl/	nip	/nɪp/	tooth	/tuθ/
mug	/mʌg/	yen	/yɛn/	near	/nɪr/	putsch	/pʊʧ/
muck	/mʌk/	yes	/yɛs/	knit	/nɪt/	pull	/pʊl/
mull	/mʌl/	yet	/yɛt/	robe	/rob/	poor	/pʊr/
mum	/mʌm/	leech	/liʧ/	roach	/roʧ/	push	/pʊʃ/
muss	/mʌs/	leaf	/lif/	road	/rod/	put	/pʊt/
mush	/mʌʃ/						

Table A.3: Words with same CV pairs

bass	/bæs/	sot	/sɑt/	leak	/lik/	more	/mor/
gas	/gæs/	shot	/ʃɑt/	meek	/mik/	pore	/por/
lass	/læs/	tot	/tɑt/	peak	/pik/	roar	/ror/
mass	/mæs/	watt	/wɑt/	pique	/pik/	sore	/sor/
pass	/pæs/	yacht	/yɑt/	reek	/rik/	shore	/ʃor/
wrasse	/ræs/	dawn	/dɔn/	seek	/sik/	tore	/tor/
bung	/bʌŋ/	fawn	/fɔn/	chic	/ʃik/	wore	/wor/
dung	/dʌŋ/	gone	/gɔn/	teak	/tik/	yore	/yor/
hung	/hʌŋ/	lawn	/lɔn/	week	/wik/	hearse	/hɝs/
lung	/lʌŋ/	pawn	/pɔn/	bitch	/bɪʧ/	curse	/kɝs/
rung	/rʌŋ/	yawn	/yɔn/	ditch	/dɪʧ/	nurse	/nɝs/
sung	/sʌŋ/	bed	/bɛd/	hitch	/hɪʧ/	purse	/pɝs/
tongue	/tʌŋ/	dead	/dɛd/	niche	/nɪʧ/	terse	/tɝs/
young	/yʌŋ/	fed	/fɛd/	pitch	/pɪʧ/	verse	/vɝs/
baht	/bɑt/	head	/hɛd/	rich	/rɪʧ/	worse	/wɝs/
dot	/dɑt/	led	/lɛd/	witch	/wɪʧ/	boom	/bum/
got	/gɑt/	red	/rɛd/	bore	/bor/	doom	/dum/
hot	/hɑt/	said	/sɛd/	chore	/ʧor/	whom	/hum/
jot	/ʤɑt/	shed	/ʃɛd/	for	/for/	loom	/lum/
cot	/kɑt/	wed	/wɛd/	gore	/gor/	room	/rum/
lot	/lɑt/	zed	/zɛd/	whore	/hor/	tomb	/tum/
not	/nɑt/	beak	/bik/	core	/kor/	womb	/wum/
pot	/pɑt/	cheek	/ʧik/	lore	/lor/	zoom	/zum/
rot	/rɑt/						

Table A.4: Words with same VC pairs

Appendix B

Symbols used in Figures

The phonetic symbol font of vowels could not be used in the figures. So the phonetic symbols were written in a different alphabet. The alphabet corresponding to each phonetic symbol is tabulated below.

Symbols in Figures	i	I	E	ae	u	o	U	c	^	a	R
Phonetic Symbols	i	ɪ	ɛ	æ	u	o	ʊ	ɔ	ʌ	ɑ	ɝ

Table B.1: Table of Phonetic Symbols used in Figures

Bibliography

[1] Christer Gobl Ailbhe Ni Chasaide. Contextual variation of the vowel voice source as a function of adjacent consonants. *Language and Speech*, 36(2, 3):303–330, 1993.

[2] Harold L. Barney Gordon E. Peterson. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184, March 1952.

[3] James M. Hillenbrand. Identification of resynthesized /hvd/ utterances: Effects of formant contour. *The Journal of the Acoustical Society of America*, 105(6):3509–3523, June 1999.

[4] Thomas R. Edman James J. Jenkins, Winifred Strange. Identification of vowels in vowelless syllables. *Perception and Psychophysics*, 34(5):441–450, 1983.

[5] Michael J. Clark James M. Hillenbrand, Laura A. Getty. Acoustic characteristics of american english vowels. *The Journal of the Acoustical Society of America*, 97(5):3099–3111, June 1995.

[6] Terrance M. Nearey James M. Hillenbrand, Michael J. Clark. Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America*, 109(2):748–763, Feb 2001.

[7] Arthur S. House Kenneth N. Stevens. Perturbation of vowel articulations by consonantal context: An acoustical study. *The Journal of Speech and Hearing Research*, 6(2):76–93, June 1963.

[8] J. L. McClelland and J. L. Elman. Interactive processes in speech perception: The trace model. In J. L. McClelland, D. E. Rumelhart, others, and eder, editors, *Parallel Distributed Processing: Volume 2: Psychological and Biological Models*, pages 58–121. MIT Press, Cambridge, MA, 1986.

[9] Adrienne Prahler. Analysis and synthesis of the americal english lateral consonant. Master's thesis, MIT, 1998.

[10] James Glass Rolf Carlson. Vowel classification based on analysis-by-synthesis. *Proceedings of International Conference on Spoken Language Processing*, 1:575–578, October 1992.

[11] Kenneth N. Stevens. *Aucoustic Phonetics.* The MIT Press, 1998.

[12] Kenneth N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4):1872–1891, April 2002.

[13] H. Thompson. Work recognition: A paradigm case in computational (psycho-)linguistics. *Proceedings of the Sisth Annual Meeting of the Cognitive Science Society*, 1984.

[14] Hsiao-wuen Hon Xuedong Huang, Alex Acero. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Developement.* Prentica Hall PTR, 2001.