Analysis and Interpretation of Glide Characteristics in Pursuit of an Algorithm for Recognition

by

Walter Sun

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

November 1996
[February 1997]

Author ....
Department of Electrical Engineering and Computer Science
November 6, 1996

Certified by ....
Kenneth Noble Stevens
LeBel Professor of Electrical Engineering
Thesis Supervisor

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

Analysis and Interpretation of Glide Characteristics in Pursuit of an Algorithm for Recognition

by

Walter Sun

Submitted to the Department of Electrical Engineering and Computer Science
on November 6, 1996, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

In the system being developed for accessing lexical items from the speech signal, an initial step is to identify landmarks which indicate the presence of vowel, glide, and consonant segments. This thesis addresses the problem of identifying the glide landmarks. The acoustic properties measured to locate glide landmarks are based on a model of glide production. This model states that a tight constriction is formed within the vocal tract, and this narrowing causes the energy in the output speech signal to be attenuated [Bickley and Stevens, 1986]. This tight constriction causes the RMS amplitude to decrease, in addition to F_1 values being very low and F_2 values being either extremely high (for /j/) or extremely low (for /w/).

The acoustic properties measured include RMS amplitude, formant frequencies, and the transitions of these properties into and out of the glide. Previous work has located abrupt consonantal landmarks [Liu, 1995]. Further, one can expect that a reasonably error-free method can be developed for locating vowels within the speech stream. Given this information, the algorithm devised in this thesis determines whether or not a glide occurs between a consonant landmark and the following vowel by determining if the acoustic properties listed above fall into a certain range. In particular, one obtains means and covariances of glides and non-glides for the given acoustic properties from a training set, and performs hypothesis testing to determine if a given measurement in the test set is a glide or not. The thesis discusses the algorithm and the results, along with the effects of context and variability in speech and how it affects the recognition process. The overall recognition results were 88.0% for glide detection, and 90.6% for non-glide detection.

Thesis Supervisor: Kenneth Noble Stevens
Title: LeBel Professor of Electrical Engineering

Dedication:

To Mom and Dad

Acknowledgments

First and foremost, I would like to express the utmost gratitude to my research advisor Kenneth N. Stevens for his guidance, assistance, and supervision of this thesis. Professor Stevens always went out of his way to assist me in finding references and giving me new ideas. Ken was more than just a research advisor to me. He showed much personal concern for me, which made working for him a pleasure. He showed faith in me by allowing me to join his research group for graduate work upon my arrival at MIT.

I want to thank the Department of Defense (Army Research Office) for their fellowship support in my graduate work. Having the National Defense Science and Engineering Fellowship has allowed me to focus full-time on graduate work, and not on ways to financially support myself.

Many thanks to ARI, CB, DW, JW, KS, MJ, and SSH for being subjects for my speech corpus. I also want to acknowledge Stefanie Shattuck-Hufnagel for her suggestions regarding the analysis of vowel stresses, and their effect on the preceding glides. I thank Sharlene Liu for helping me discuss work in her thesis, and her tips and thoughts regarding my thesis. Likewise, Mark Johnson deserves much praise for all of his help and assistance. Despite being very busy and having the same thesis deadline as myself, Mark never hesitated to address any of my questions, from software to hardware to theoretical to course-related issues.

Speaking of software/hardware support, I thank Seth Hall for the time he spent trying to get everything working in the computer lab, especially in the weeks while I was writing my thesis. Many things went awry, but Seth put all of them back in order, and he was willing to address these problems one by one.

In addition, I want to thank Andrew Kim for his help in allowing me to transfer much of my work onto Athena, along with using his machine Samoyed. I also thank him for being a person I could talk to regarding technical issues, both with coursework, research, and any other topics of discussion.

Thanks also go to Dewey Tucker, my roommate, for putting up with my odd

hours at times, and for being another person to talk to regarding my research and coursework. Likewise, I thank Nick Laneman for helping to keep me sane by getting me to go out and sail, play basketball, and other sorts of outdoor activities.

Many thanks go to Judy Chang for her constant support and care towards my progress on this thesis. Judy was very encouraging, and always available for me to talk to in the best and worst of times.

Thanks go to Sharon Manuel, Jeff Kuo, Elizabeth Choi, Helen Hanson, Marilyn Chen, Arlene Wint, Phil Lin, Mike Lii, Jerry Chen, and Todd Ulmer for their technical expertise and moral support. Much appreciation goes to the Oral Qualifying Exam committee study group (Andrew, Dewey, Nick, and Edmund Yeh) for the knowledge we've bounced back and forth over the summer in preparation for our upcoming oral exam.

I want to also thank all of my friends at Boston Chinese Evangelical Church (BCEC), those at MIT's Chinese Christian Fellowship, and our graduate small group. They have been my family away from home, and the moral support I have received from them has made a big difference in my adjustment to graduate school at MIT. I thank God for bringing me to Boston and finding BCEC, along with allowing me to know Him better, and for looking after me in the best and the worst of times.

Finally, I want to thank my family, my mother, father, brother Ray, sister Nancy, and brother-in-law Greg for their support of my endeavors. I thank my parents for the sacrifices they made to make my college education and future possible, and for the great upbringing that they provided me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the process of speech recognition, a computer must recognize and ultimately interpret the utterances of human beings, much in the way two human beings interpret and understand each other. Two major techniques of speech recognition can be found in current research. In one method, a speaker trains a recognition device by speaking sample utterances into the machine before the actual recognition process begins. This method usually relies on the idea of a hidden Markov model or a neural network. The computer tries to "learn" the particular speaker and his/her variabilities through the training data, and then uses this information to help "understand" what the speaker says. Another major technique in speech recognition is the knowledge-based approach. Here, algorithms are developed based on the knowledge of the speech perception process and the acoustics and physiological aspects of speech production. From this information, one tries to model the production of each sound. Then, during the process of speech recognition, the computer tries to match every produced sound to one particular model from all of the possible models involved in speech production for the particular language of interest. Many speech recognition systems incorporate aspects of both techniques.

The thesis research discussed here focuses on the approach of knowledge-based analyses. A knowledge-based approach seeks to model what a human being sees and hears. In the area of speech recognition, this is accomplished by determining how differences in vocal tract shape and how glottal excitations influence the acoustic waveforms measured in sound recordings. In particular, different phonemes, characteristic sounds in a language, each have unique features that can be identified, although the acoustic properties depend to some extent on the context. These properties are present in the sound in the vicinity of landmarks that are located at a release or closure for consonants and at a minimum or maximum vocal tract opening for glides and vowels. Upon completion of this process, a labeled waveform can be translated into a meaningful sequence of words by accessing a lexicon.

In this thesis, a knowledge-based approach towards detecting and recognizing the glides (/w/ and /j/) is developed. Detecting glides is one aspect of a complete lexical access project, which already has in place a procedure for locating landmarks for abrupt consonants [Liu, 1995]. Finding glides is somewhat difficult because, unlike abrupt consonants which have sudden changes in energy, movement into and out of a glide is generally gradual, so measured acoustic parameters have smooth transitions into and out of the glide.

Previous work in this area of recognition was performed by Sharlene Liu [1995] and Carol Espy-Wilson [1987]. Liu's work focused on finding abrupt consonant landmarks, and her findings are used to aid in the recognition of glides discussed in this thesis. Espy-Wilson's work focuses on recognizing semivowels, both glides and liquids, in a controlled environment using a carrier phrase, and distinguishing between the different semivowels.

Additional goals of this thesis are to examine variabilities between individual glides and within phonemes and to devise models that might account for such interspeaker and intraspeaker differences. Recognition was performed on both isolated and continuous speech. A subset of the continuous speech was used for training the data, while the remainder of the speech served as the test set.

Chapter two provides a background about glides from previous work performed by others, discusses landmark detection, and introduces the classification method of hypothesis testing, the method used in this thesis for glide recognition. Chapter three discusses the procedures followed to create the corpus of data, generate the waveforms, and extract the features from the data. Chapter four summarizes the different preliminary methods considered, and reasons why these ideas did not work or were not pursued further. Chapter five provides an in-depth discussion of the results obtained from the VCV database, and explains how different vowel contexts affect recognition. Chapter six discusses the final results obtained both in the VCV and sentences databases. Finally, the seventh chapter gives a summary and provides suggestions for future work.

Chapter 2

Background

Glides have been looked at in the past by several authors, including [Lehiste and Peterson, 1961], [Bickley and Stevens, 1986], [Espy-Wilson, 1987], [Arman and Stevens, 1993], and [Stevens, forthcoming]. The glide detection algorithm being developed here fits into a larger lexical access project, which has the ultimate goal of accessing words in a lexicon by making appropriate measurements to identify the distinctive features. As part of this project, Liu [1995] completed a thesis which discussed the procedure for finding acoustic landmarks of abrupt consonants. Discussion of this past work on glides and landmark detection is summarized in this chapter.

2.1 Characteristics of Glides

Glides belong to the class of sonorant consonants. This class of consonants includes nasals and liquids, as well as glides. Sonorants usually have steady, periodic voicing at the glottis. They usually have extreme formant values and they often have extra pole/zero pairs in the vocal tract transfer function. The extra pole/zero pair, as seen in a nasal, comes from the fact that the nasal cavity is open, leading to a new passage in the vocal tract. Stevens [forthcoming] goes into greater detail with regard to the characteristics of glides.

Glides are members of the class of segments known as semivowels. This name is given to glides because, like vowels, all of the formants are excited in the vocal tract.

In essence, the vocal tract is in a vowel-like configuration, with the /w/ like the /uw/, and the /j/ like the /iy/, except having a tighter constriction [Arman and Stevens, 1993]. Furthermore, glides are always in onset position in a syllable, and hence are followed by vowels. Glides, as their name may indicate, provide a smooth transition of formant frequencies and amplitudes into the following vowel.

2.1.1 Amplitude and Spectrum Shape for Glides

Glides are generally voiced, although cases occasionally occur where they are produced without glottal vibration. This circumstance can occur only when the segment preceding the glide is an unvoiced consonant. The word *twin* is an example of such an exception. The overall amplitude in a glide is usually decreased relative to the amplitude in the following vowel. Bickley and Stevens [1986] and Stevens [forthcoming] report that the amplitude of the first formant peak at the most constricted point in a glide is at least 2-3 dB lower than that of vowels, with decreases becoming as large as 7-8 dB when the adjacent vowels are low vowels having high F_1. The amplitude of the second formant peak is also lower by about 9 dB than the amplitude that would be observed with a vowel having about the same formant frequencies. These observations suggest that vocal-cord vibrations can be altered slightly by a tight constriction, with the effect being a change in the amplitude, shape, and area of the glottal pulses.

2.1.2 Formant Frequencies

The formant locations for glides are at "extreme" values. F_1 drops to about 260 Hz on average for males, and a little higher for females. This value can be predicted from the Helmholtz resonance corrected for the mass of the walls. For /j/, the back of the vocal tract is a volume which is terminated with a narrow opening in the oral cavity. Thus, it can be approximated by a Helmholtz resonator. For /w/, the narrow opening is at the lips with another constriction formed by the raised tongue body. This lowest natural frequency is computed using the following formula:

$$F_1 = \sqrt{(F_1')^2 + (F_{1_c})^2}, F_1' = \frac{c}{2\pi\sqrt{\frac{Vl_c}{A_c}}} \tag{2.1}$$

where F_{1_c} is the natural frequency due to the mass of the vocal tract walls and the compliance of the air volume and F_1' is the Helmholtz resonance. For /j/, the volume behind the narrow constriction is about 50 cm^3, with the length of the constriction approximately 3 cm, and the cross-sectional area around 0.17 cm^2. The speed of sound in the vocal tract is approximately 35400 $\frac{cm}{sec}$. These values yield an F_1' of about 190 Hz. The natural frequency from the wall effects (mass and compliance) result in an F_{1_c} of about 180 Hz, leading to a value of 262 Hz for F_1. The presence of the wall effects reduces the variability of F_1 in glides due to anatomical differences in vocal tract length and variation in the size of the constriction.

F_2, likewise, is at an extreme. For /w/, F_2 is in an extremely low position. This is due to the secondary constriction caused by the raising of the tongue body in the velar region. The decrease in area near the middle of the vocal tract, a region where the volume velocity is a maximum for F_2 frequencies, causes a decrease in the formant frequency. This effect can be shown acoustically through an analysis of perturbation theory.

On the other hand, F_2 for /j/ is at an extremely high position. This can be attributed to the raising of the tongue blade in the palate, a region in the vocal tract where the volume velocity is a minimum for F_2 frequencies. Hence, with perturbation theory, one would expect an increase in F_2.

Espy-Wilson [1987] made measurements of glide formant locations in word-initial (#GV), prevocalic (CGV), and intervocalic (VGV) position, where # = word boundary, G = glide, V = vowel, and C = consonant. Although the measured values for F_1 were not as low as the theoretical value of 260 Hz, they were still low. Table 2.1 shows the results obtained by Espy-Wilson. These values also show the high F_2 for /j/ and the low F_2 for /w/.

The formant frequencies at the point of maximum constriction for glides are ex-

Glide	F_1 (Hz)	F_2 (Hz)
Word-initial		
/w/	347	739
/j/	281	2190
Prevocalic		
/w/	351	793
/j/	305	2190
Intervocalic		
/w/	349	771
/j/	361	2270

Table 2.1: Formant values for glides as measured by Espy-Wilson [1987].

pected to depend on the gender of the speaker, much like the formant frequencies for vowels depend on gender. These differences are a consequence of a difference in average vocal-tract length for males and females, which is about 17%. However, based on data for the vowels /iy/ and /uw/, it is expected that gender differences in F_1 for /w/ and /j/ and in F_2 for /w/ are much smaller than 17%, whereas differences in F_2 for /j/ might be as large as 20% [Peterson and Barney, 1952; Fant, 1959].

2.1.3 Rates of Formant Movements for Glides and Non-Glides

Previous work has shown that formant transition times for stops are much shorter than those for glides. This result, and its corresponding relation to the transition times for RMS amplitude, will prove useful within this thesis. Following is a summary of past studies on these transition times.

Comparison of Initial Transition Times between Stops and Glides

Miller and Baer [1983] analyzed the durations of initial formant transitions for both /b/ and /w/ to show that this acoustic property could differentiate these two consonants. Miller and Baer considered two variables in their analysis: the duration of the F_1 transition and the length of the syllable (/ba/ or /wa/) containing this transition. The data showed that while the transition time for /b/ in /ba/ stayed more or less constant at about 40 milliseconds with increasing syllable duration, the transition time for /w/ in /wa/ increased linearly with syllable duration. The transition duration for /wa/ varied from about 50 to 150 milliseconds. Consequently, for longer syllable durations, larger separation of transition times could be seen between /b/ and /w/. In fact, for all syllable durations studied (from 100 to 700 milliseconds), the mean formant transition time for /w/ was longer than that for /b/. However, for syllable durations less than 250 milliseconds, some overlap of data was found.

Thus, from Miller and Baer, one can conclude that longer formant transition durations are expected for the glide /w/ than for the stop /b/. However, since some overlap exists in these measurements for quickly spoken syllables, one can expect occasional errors when using this as an absolute rule.

Duration of Initial Transitions from Consonant into Vowel

Lehiste and Peterson [1961] discussed formant movements related to transitions, glides, and diphthongs in their paper. In their work, they reported measurements of the duration of formant transitions and rates of change of formant movements. Lehiste and Peterson found that labials have shorter initial transitions than lingual consonants. The shorter transitions for labials are presumably a consequence of the fact that in labials, the tongue body is free to prepare itself for the position of the following vowel. Consequently, only a small tongue body movement is required, resulting in a shorter transition time. The importance of this result is that labials have much shorter durations than lingual consonants, glides included. Mack and Blumstein [1983] show yet further support in the following section.

More Transition Contrasts between the Stop and the Glide

The duration of the F_1 transition for /b/ and /w/ with different following vowels was also measured by Mack and Blumstein [1983]. Their data also showed significantly longer F_1 transitions for the glide. Mack and Blumstein then hypothesized that, based on the nature of the articulatory configuration for stops and glides, one would expect similar rates of change for energy measurements. Their experiments confirmed this, demonstrating that the rate of change of energy at a stop consonant release was significantly greater than the rate of change for a glide.

Formant Transitions and their Effect on Amplitude Transitions

As briefly discussed in Mack and Blumstein above, stops have much faster rates of change of energy than glides. From an articulatory standpoint, one can observe that the release of a closure in the vocal tract for stops leads to an abrupt increase in transglottal pressure, which causes a sudden increase in the RMS amplitude of the signal. In glides, no such pressure build-up occurs. Consequently, such a sudden change would not be expected to occur.

In relation to formant frequencies, a sudden increase in F_1 with all other factors held constant will cause a sudden increase in the overall RMS amplitude of the signal because the amplitude of the formant peak in the transfer function is proportional to the either the frequency or the square of the frequency of the formant at low frequencies. This is true because the amplitude of the first formant component of the vocal tract transfer function is equal to the first formant frequency divided by its bandwidth. The spectrum amplitude of the glottal waveform at low frequencies is approximately constant or decreases only as the inverse of the frequency. In addition, the amplitude of the radiation characteristic of the signal from the mouth opening is proportional to the frequency of the signal. Thus, the observed amplitude of F_1 in the speech signal, which is determined by a cascade of the vocal tract transfer function with the glottal waveform and the radiation characteristic (a product of the amplitudes in the frequency domain), will be proportional to F_1 or to the square of F_1 divided by the bandwidth. For the case being discussed here, the bandwidth is assumed to remain constant. Consequently, the overall energy, which is due in large part to the low frequency energy, will increase when F_1 increases.

2.2 Previous Work on Lexical Access

Previous studies in the Speech Communication Group have focused on automatic semivowel detection in a controlled environment, and landmark detection of abrupt consonants in continuous speech. Results found from such studies are discussed below.

2.2.1 Semivowel Detection

Espy-Wilson [1987] used an acoustic-phonetic approach to arrive at the recognition of semivowels in the controlled environment of a carrier phrase ("— pa"), where the — is a polysyllabic word containing a semivowel. The "pa" at the end reduces the possibility of glottalization that often occurs in utterance-final position. Recognition was limited to voiced and nonsyllabic semivowels. The work found in the following chapters furthers Espy-Wilson's work by performing glide recognition within continuous speech for both voiced and unvoiced glides, and for some glides within words in utterance-final position. The analysis of Espy-Wilson involved (1) specifying the features necessary to distinguish the semivowels, (2) finding ways to account for intraspeaker and interspeaker variations, (3) extracting these features from the speech waveform, and (4) combining the acoustic properties to create an algorithm for recognition.

The detection of semivowels required finding minima and maxima in the F_2 and F_3 formant tracks. After these landmarks were located, different acoustic properties were considered to see if this region was possibly a glide or a liquid. A landmark having a low F_1, a low F_2, and gradual onset time into the vowel (the onset time was defined as the time between the maximum change of energy going into the vowel and the energy peak within the vowel) was classified as a /w/. One having low F_1, high F_2, and gradual onset time into the vowel was determined to be a /j/.

In the end, Espy-Wilson found a range of recognition rates depending on the context. For /w/, she found recognition ranges from 21% (intervocalic) to 80% (word-initial); for /j/, between 78.5% and 93.7%. Overall, the recognition rate for semivowels including both liquids and glides ranged between 87% and 95%.

Band	Frequency Range (kHz)
1	0.0 - 0.4
2	0.8 - 1.5
3	1.2 - 2.0
4	2.0 - 3.5
5	3.5 - 5.0
6	5.0 - 8.0

Table 2.2: Frequency Bands Used by Liu

2.3 Landmark Detection

Liu [1995] completed work on landmark detection of segments with abrupt acoustic changes. Liu used the Lexical Access from Features (LAFF) database, a database with few consonant clusters or syllables, the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, noisy speech (LAFF database corrupted with noise), and telephone speech (from network TIMIT, a database of information similar to TIMIT, but with the speech transmitted through telephone lines in New England across NYNEX). In her landmark detection algorithms, Liu separated the spectral signal into six frequency bands. Table 2.2 summarizes these frequency bands. The frequency bands were chosen so that each band could track the strength of particular formant frequencies.

Next, six different possible labels were appended to the data (+g, -g, +s, -s, +b, -b). The landmark most useful for the analysis made on glides is the +g landmark. The g stands for glottis, and a +g landmark indicates the point where strong glottal vibration, or voicing, begins. These landmarks are found when the rate of change of band one energy is a maximum. Likewise, the -g landmark (end of voicing) is determined by maximum negative rates of change of band one energy. The s landmark indicates an abrupt sonorant (nasals and /l/). These landmarks are located by searching for peaks in the rates of change of band energies two through five. Possible s landmarks must lie between +g and -g landmarks because of the assumption that these abrupt sonorants are voiced. The sonorant detector is examined briefly in the analysis in chapter five. Finally, the b landmark is used to locate bursts. Similar to the sonorant landmark, the process of locating bursts involves finding extreme rates of change in the higher order energy bands (bands three through six). As opposed to the sonorant landmark, possible regions for bursts must be unvoiced. Thus, a burst landmark must not lie between a +g and a -g. This landmark detector was successful in clear speech, having a detection rate of 90%. Insertion rates ranged from 6% to 9%.

2.3.1 Executing the Landmark Detector on Speech

Running the program to obtain landmarks required three steps. First, spectrograms of the data were generated using *sgram* of ESPS. The spectrograms were 512-point FFTs of six millisecond windows, updated every millisecond. Since the sampling rate was 16 kHz, the FFT size used allowed for a frequency resolution of 32 Hz. One spectrogram was pre-emphasized with a first difference (1-z^{-1}), while the other was not pre-emphasized. Next, a program entitled *hpror* was executed to create rate of rise data to be used in the landmark routine. Finally, another program called *lm* was used to locate the abrupt landmarks. The C code that was made may be found in the Appendix of Liu's thesis.

2.4 Hypothesis Testing

The method of hypothesis testing was the method used in this thesis to determine if a given set of measurements qualified as a glide. Binary hypothesis testing is appropriate for glide detection because the hypothesis one wants to test is whether or not a glide is centered at a given point in time (only two possibilities). The means and covariance matrices used for the analysis were determined from a training set of data. Following will be a brief explanation based on information obtained from Willsky et al. [forthcoming] regarding the concept of hypothesis testing. A much more complete discussion may be found in Willsky et al.

Based on a Bayesian approach, hypothesis testing involves knowing the apriori probabilities of each of the two different events, along with modeling the probability density functions of each of the events. Since the exact probabilities of each of the two events were not known, it was assumed that the two possibilities (glide versus non-glide) were equally likely. Normally, this may seem unusual, but in the analysis given in the following chapters, it is noted that the test for a glide occurs by taking data starting thirty milliseconds from the point of the onset of voicing. In cases where the distance between the landmark and the following vowel was less than thirty milliseconds, the region in question was automatically labeled as a non-glide section. This is reasonable because the duration from a glide into a vowel is more than thirty milliseconds, with minimums being only as small as fifty milliseconds in continuous speech. Thus, only regions where the duration from landmark to vowel was more than thirty milliseconds were considered in the hypothesis test. The remaining speech segments in question were modeled to have a 50% chance of being a glide, and 50% chance of being a non-glide. Albeit a rough estimate, setting the two apriori probabilities equal served as a reasonable assumption which would at the same time simplify computation. Next, the probability distribution had to be modeled. The variability of the parameters selected was modeled to be distributed in a Gaussian probability density function around its mean. This distribution was used as a simplification for detection without specification of context, and can likely be improved if specific contexts are considered separately since the measurements will be localized around different means for different vowels which follow the glides. However, in this thesis, a general hypothesis test is performed on each new set of measurements without regard to specific context. The covariances of this distribution were determined by taking the covariance of the sampled data from the training set. Likewise, the means were the sample means of the training data.

The hypothesis test compares the probability of one hypothesis given the observation versus that of the other hypothesis. The one which is greater is the one more likely to be the correct hypothesis. As a result, this hypothesis is taken as the prediction. Stated in the form of equations, with Y being the observation vector, p(x) being the probability of some event x, H_0 being one hypothesis, and H_1 being the other,

H_1 is chosen when

$$p(H_1 \mid Y) \geq p(H_0 \mid Y) \tag{2.2}$$

H_0 is chosen when

$$p(H_1 \mid Y) < p(H_0 \mid Y) \tag{2.3}$$

Applying Bayes' Law, which says that

$$p(b \mid a) = \frac{p(b)p(a \mid b)}{p(a)} \tag{2.4}$$

provided that p(a) $\neq$ 0, and the stated assumption that p(H_1) = p(H_0) = $\frac{1}{2}$, one obtains

$$p(Y \mid H_1) \bigcirc p(Y \mid H_0) \tag{2.5}$$

where one would select H_1 if $\bigcirc$ were $\geq$, else H_0 if $\bigcirc$ were $<$.

In this thesis, H_1 served as the hypothesis of a set of observations being consistent with a glide, while H_0 is that consistent with a non-glide. The variables involved in vector Y are RMS maximum slope, RMS amplitude range, F_1 maximum slope, and F_1 frequency range (these measurements will be discussed later). Since the measurements are not independent, a covariance matrix (C) had to be used to represent the variances in the measurements (and not four independent variances). In addition, the means of each of the measures had to be determined. The covariance and means were determined from samples in the training data set. With this information, the probability density function p(Y | H_i) (where i could be zero or one) was constructed as follows:

$$p(Y \mid H_i) = \frac{1}{(2\pi)^2\sqrt{det(C)}} e^{(-\frac{1}{2}(Y-m_Y)^T C^{-1}(Y-m_Y))} \tag{2.6}$$

for i=0,1, where m_Y is the sample mean of the observation vector Y, and det(C) is the determinant of C. This density function was used to compute the probabilities of each observation for i=0,1, and then classification was made based on the hypothesis test.

Chapter 3

Development of Database

Two different databases were used in the thesis, one being isolated vowel-consonant-vowel utterances, and the other being continuous speech. This chapter discusses the characteristics of the former and the development of the latter. The two databases allowed one to compare the characteristics of the glides under different phrasal contexts.

3.1 VCV Database

The first major step in the process of developing methods for detecting and recognizing glides was to find and/or create a database with which to do analysis. An existing vowel-consonant-vowel (VCV) database was available for use as training data for the algorithm being developed. The VCV database was good for initial measurements because it was a well-controlled set of clearly spoken utterances. The VCV database is a set of vowel-consonant-vowel utterances created within the Speech Communication Group at MIT. This database has three speakers, two male (DW and KS) and one female (CB). Each speaker produced all possible combinations of the vowel-consonant-vowel utterances for the vowels /aa/, /ah/, /eh/, /iy/, /ow/, and /uw/ and the consonants /b/, /ch/, /d/, /dh/, /dj/, /f/, /g/, /h/, /j/, /k/, /l/, /m/, /n/, /ng/, /p/, /r/, /s/, /sh/, /t/, /th/, /v/, /w/, /z/, and /zh/.

To extract the data from the VCVs, the consonant and the final vowel were hand-labeled through listening to the utterances and visually observing the spectrograms. The final vowel was labeled at the maximum of RMS amplitude, while the consonant was labeled at the onset of voicing for the obstruents, and at the appropriate energy minimum for the sonorants. The region between the consonant and the final vowel was examined to determine transition rate and ranges, as seen in chapter 5.

After using the VCVs as a training database, another set of speech waveforms was needed for the testing. To verify the robustness of the algorithm developed, a database which had more variability and a variety of contexts was desired. Thus, a set of sentences was created by careful construction. The next section provides details on how this sentences database was created.

3.2 Sentences Database

The sentences database consisted of forty-two sentences spoken by five native English speakers. As will be explained at the end of this section, two of the forty sentences had to be removed. The sentences were created with several goals in mind. A primary goal was that the sentences had to be rich in glides. Another goal was to design utterances in which the glides occurred in a variety of different contexts. Words were found in which glides were followed by a variety of different vowels. A set of words containing glides followed by a number of following vowels was developed by referring to the list of one-syllable words in Moser [1957]. All of the following possible vowel sounds were sought for words containing glides. After a list of possible words was created, sentences containing these words were made. Words containing glides were located at the beginning, in the middle, and at the end of sentences to allow the most possible variability. Some of the glides were also located before reduced vowels to see the effects of vowel reduction on detecting glides. In addition, some glides were placed in intervocalic position, while others were located after consonants. The remainder were located at the beginning of a new syllable. For glides located at the beginning of a syllable, some were sentence-initial, while others were in the form of either C#GV or V#GV. Of the forty sentences eventually used for analysis, twenty had glides in the initial word (fourteen of these glides were the first phoneme in the word, also) and twenty-three had glides in the last word of the sentence. In total, the forty sentences had 128 glide occurrences, with one or two less for some speakers because of variability between different speakers. That is, some speakers did not produce glides in all places where glides can be produced in some dialects. For example, the word *tune* can be pronounced as /tun/ or /tjun/. Twelve of the 128 glides were followed by reduced vowels, while the remainder were followed by full vowels. Table 3.1 shows the list of sentences in the sentences database. Each utterance was spoken twice by each speaker (the characteristics of the speakers will be explained in the following section), and the first repetition was taken, unless a mistake was made, in which case the second recording was used.

It should be noted that two sentences that were recorded were removed from the database because of errors in their production. These were the sentences *Blow away your woes, and you will quickly recover,* and *Will you put another yule on the fireplace?*

3.3 Recording and Labeling Procedures

Five adult, native English speakers were chosen for recording the sentences. Three subjects were male (ARI, KS, and MJ), while the other two were female (JW and SSH). None of the speakers had any noticeable accents, and all but one (ARI) had previous experience in making recordings for data analysis. The four who had experience were subjects in a Lexical Access from Features (LAFF) database, which was used extensively by Liu [1995] in her work on lexical access. The recordings were made in a sound-proof room. Measurements were made of the ambient noise level inside and outside the sound-proof room with a sound level meter (Quest Electronics). The sound pressure level ('A' scale) in the room was 27 dB. Sentences were spoken with the microphone about 8 inches away from the mouth of the speaker, with the omnidirectional microphone at a level a couple of inches above the mouth of the speaker to avoid the puffs of air which come about when one speaks. The signal

1	We were away a year ago.
2	You should use your yacht.
3	Be sure to reuse after you use.
4	You went to rewire items in Gwen's yacht.
5	Few twins like pure cubes of sugar.
6	Blow them, and the leaves will fly away.
7	Ask Gwen to play a tune.
8	The twins dwell in their yacht.
9	Gwen went to rewire the twin's cabins.
10	You will dwell upon the cube problem.
11	We used the yacht a year ago.
12	Blow the horn, and it will play a tune.
13	You will yell if water gets in the yacht.
14	Reuse the yarn first.
15	The young man dwelled upon his woes.
16	We wanted to yell while at the yellow yard in Watertown.
17	Which way to Yellowstone?
18	You will learn if Will walks you through it.
19	Do not dwell on yesterday's whims.
20	The willow tree wimpered in the wind.
21	The youth used a wireless phone on the yacht.
22	Yukon is in Western Canada.
23	This yacht yawed away from its way to Yorktown.
24	Gwen will quickly buy a ball of twine.
25	Swim towards your willow tree.
26	Move to your winter home.
27	Swing at the large, yellow ball with the bat.
28	Weave a sweater with the yarn.
29	Take the wheat to the market.
30	Your car wheel is wobbling.
31	The worm washed away in the water.
32	He yearns to increase his net worth to millions.
33	The wolves yanked apart the wool.
34	Are you going to the yacht race at Yale this year?
35	Egg yolks are yummy.
36	We yawn when we are tired.
37	Without warning, we yelled out to the yacht in the water.
38	The yeast has not yet risen.
39	Many legends of yore involve the city of York.
40	Newborn babies yearn to yell.

Table 3.1: Sentences database (rich in glides).

to noise ratio (SNR) was found to be around +30 dB using the sound level meter, 'A' scale.

The utterances were low-pass filtered at 7500 Hz and then digitized at a sampling rate of 16000 Hz. Spectrograms were made for each of the utterances, along with formant tracks created using ESPS [Entropic Signal Processing System, 1992] according to the algorithm set forth by Talkin [1987]. The spectrograms were individually analyzed whenever the formant values obtained automatically seemed questionable. The parameters used for the formant tracker were as follows: 40 msec cos^4x window, frame updates every 2.5 msec, preemphasis using a filter with z-transform (1-0.7z^{-1}), tracking of four formants under 5000 Hz for females, five for males. The reason why four formants were tracked for females instead of five comes from the fact that females have higher formant values than males, so one would expect that only four formants reside under the 5000 Hz rate. Although the frequency range available was from 0 to 7500 Hz (based on the low pass filter performed previous to digitizing), it was determined that formants above 5000 Hz would not be needed in any of the work performed. Consequently, in an effort to compute just what was necessary, the formant tracker looked for formants only betwenn 0 and 5000 Hz. This idea, which was suggested by Johnson [1996] (via personal communication with author), made the tracking much more accurate.

The vowel locations of the data were then hand transcribed along with the locations of the glides. A program to locate landmarks, written by Sharlene Liu [1995], was run on the sentences to find the onsets of voicing (+g landmarks) from consonants into the vowels. The landmarks, along with the vowel locations, were needed because the final algorithm searched the RMS amplitude in this region for maximum rates of change and amplitude ranges. The glides were hand labeled so that automatic recognition results could be checked against these results. An example of this labeling process is shown in Figure 3-1 for the sentence *Swim towards your willow tree* spoken by KS. The phonetic transcription of this sentence, as determined by listening to the sentence, is */s w ih m/ /t ao r d z/ /j er/ /w ih l ow/ /t r iy/*. Liu's program gave voicing (+g) landmarks at 0.31, 0.68, 0.97, and 1.53 seconds in the utterance.

Figure 3-1 shows these landmarks with tall vertical bars.

The vowels were labeled in the middle of the segment, as determined by the maximum RMS amplitude. Glides were labeled at the minimum of RMS amplitude. Nasals and liquids were also marked at their midpoints. The middles for the nasals and liquids were determined by finding the beginnings and ends, and locating the midpoint between them. Nasals and liquids were labeled because it was expected that these sonorants would occasionally be mistaken as glides in the automatic detection, and some measure of these errors was desired. The middles were chosen because it was more likely that the middle would fall in the range of analysis (from thirty milliseconds beyond the onset of voicing after an abrupt consonant to the following vowel) whenever it was mistaken for a glide. The short lollipops in Figure 3-1 locate the vowels and the sonorants.

For the RMS amplitude computation, different window sizes were analyzed. Table A.1 in the appendix shows the m-file written in MATLAB used to compute the RMS amplitude of speech signals.

Originally, smaller windows were thought to be better. However, these windows tended to yield similar results to longer windows. The only noticeable difference was that the smaller rectangular windows tend to create bigger dips in consonant regions, as expected. Since this held for both glides and non-glides, no advantage was gained with a smaller window. Eventually, the window length chosen was 49 msec. Figure 3-2 compares the RMS amplitude for the utterance *Swim towards your willow tree* by KS, using different time windows. Another window size may have led to better overall results, especially in nasals where the abruptness of closures might have been lost in the analysis with the smoothing due to the large window size. Optimization of the window length can be a topic of future work.

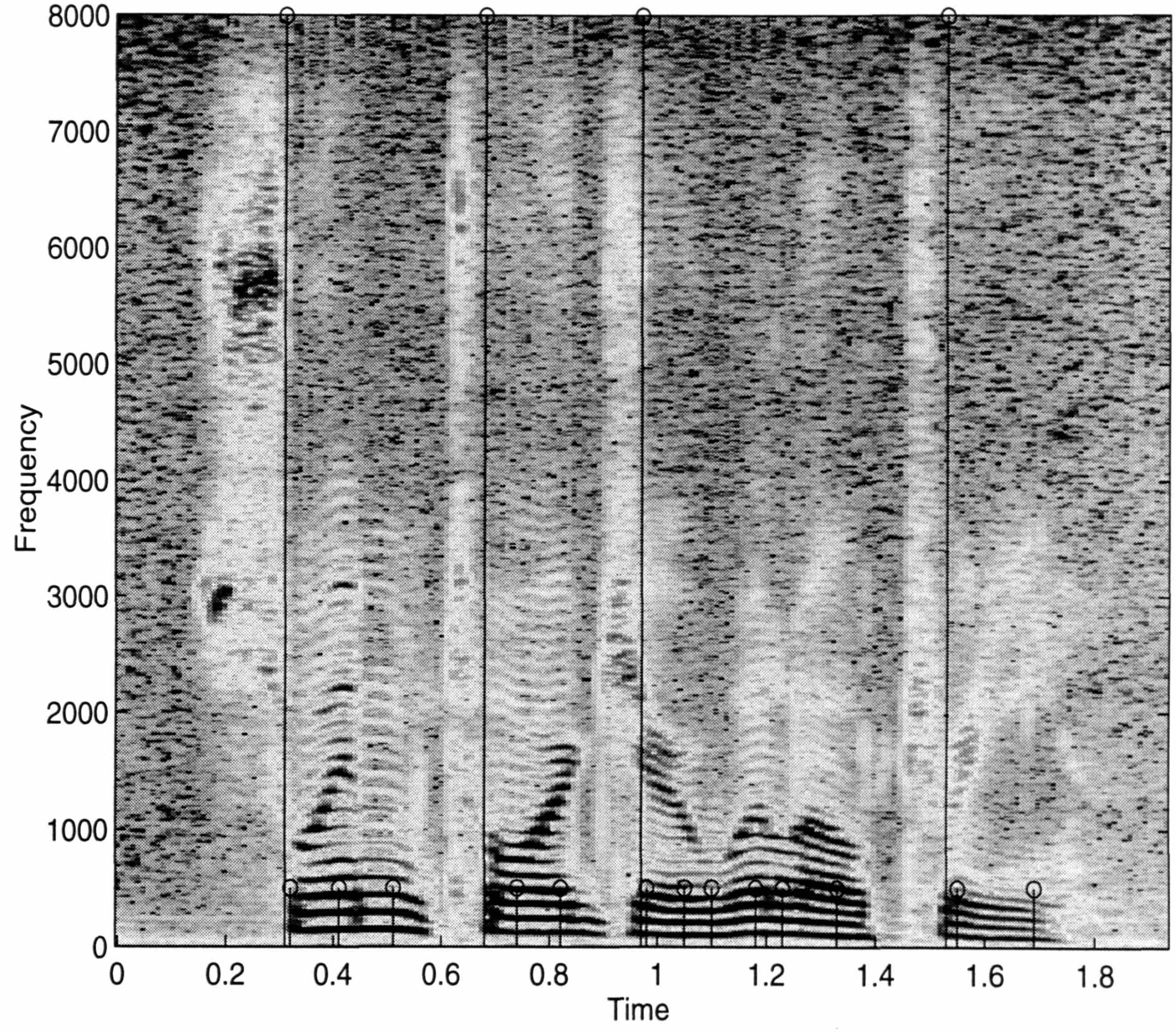

Figure 3-1: Labeled spectrogram of the utterance *Swim towards your willow tree.* The tall vertical lines indicate onsets of voicing following non-sonorant consonants, while the short lollipops show locations of vowels and sonorants.

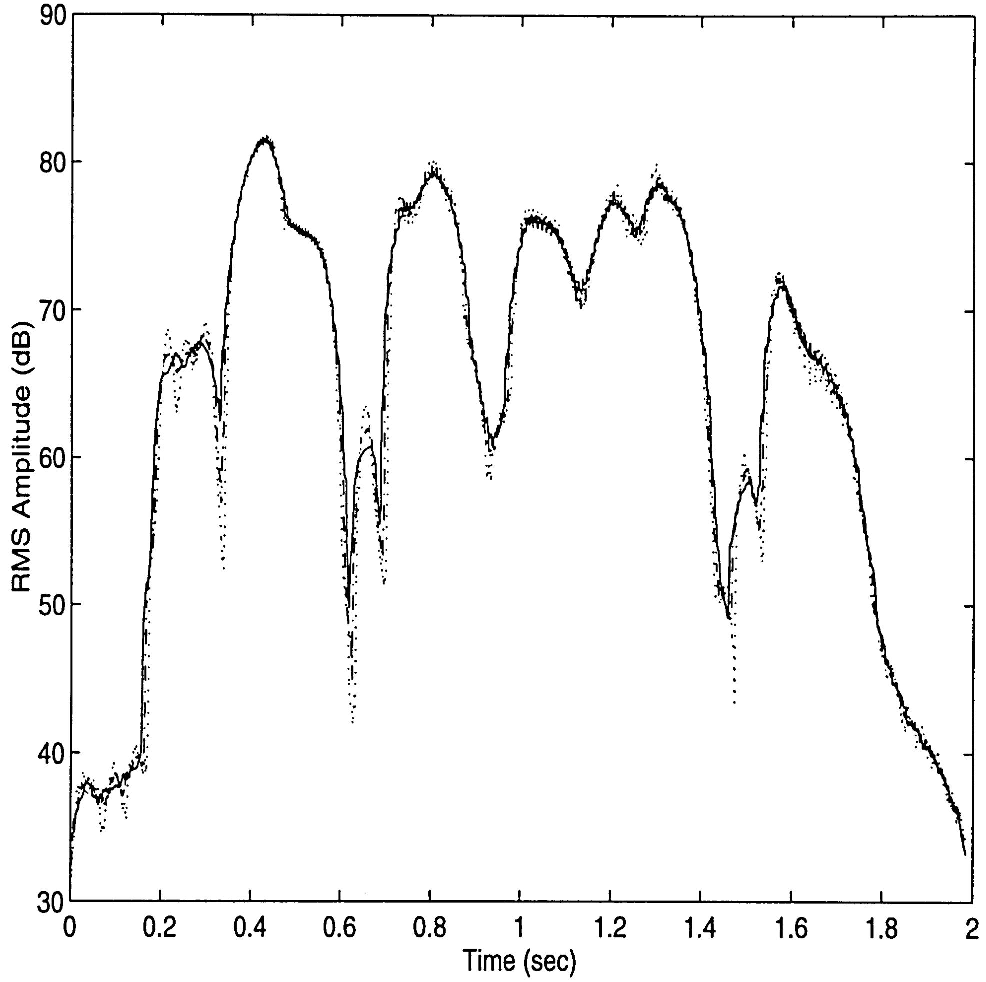

Figure 3-2: Comparison of RMS amplitude for different rectangular window lengths in *Swim towards your willow tree.* 49 msec window in a solid line, 35 msec in dash-dot, and 20 msec in a dotted line.

Chapter 4

Preliminary Experiments

The development of the final algorithm described in the next chapter evolved through a series of approaches that looked at many different features of the speech signal. Characteristics which were analyzed included the root-mean-square amplitude, the formant frequencies (primarily the first three formant values), different band energies, voicing, and the rates of changes of these values. In each of the experiments, three issues had to be considered. They were (1) what parameters or measurements should be extracted in order to detect glides, (2) how could measurements be obtained automatically without error, and (3) how should the measurements be combined for classification. Locally adaptive thresholding and best fit sinusoids, two techniques first considered but then dismissed, are discussed in this chapter. The reasons against using each of these two methods are explained. A reader interested in pursuing future work may find this chapter useful in preventing himself or herself from running into the same problems encountered here.

4.1 Detection via Formant and RMS Amplitude Thresholding

Originally, an approach using formant frequencies appeared quite promising. After all, the extreme locations of F_1 and F_2 and the smoothness of the formant movements seemed to be excellent ways to locate the glides. An algorithm was developed that used formant frequencies and RMS amplitude as factors in determining if a glide was present in a particular region in the speech waveform. The method involved applying a locally adaptive thresholding method adapted from Niblack [1986] to F_1, F_2, and the RMS amplitude of the speech waveform, and then combining these results with absolute thresholds to determine where the glides are.

4.1.1 Obtaining Measurements

First, measurements had to be performed on the data to obtain the parameters needed. The RMS amplitude was computed using a 49 millisecond window. The code used to compute this is included in Table A.1 in the appendix. Formant tracks were obtained using the formant tracking routine in Entropic Signal Processing Systems based on work developed by Talkin [1987]. To reduce the effects of noise, the data was smoothed with a ten millisecond moving averager. This duration was chosen because the finite movement rate of the tongue and the vocal cavity is slow enough that abrupt changes within ten milliseconds can be deemed to be noise.

4.1.2 Selecting Thresholds for Detection

Next, a method of automatic detection was considered. Because formant values and RMS amplitude are time-varying (different speakers have formants that differ somewhat, and RMS amplitude can vary based on how loud someone is speaking), a locally adaptive method was implemented along with the absolute measures. Figure 4-1 illustrates when a locally adaptive method works better than just absolute measures. In this example, an absolute threshold of 50 dB (dash-dotted line) allows one to find the glide at 0.25 sec, but as the overall amplitude in the waveform increases from .4 to .7 seconds, perhaps caused by an increase in speaker sound intensity, the second dip in energy is not detected at 0.775 sec using absolute measures. A locally adaptive thresholding method detects both of the dips in energy. Dotted lines indicate local thresholds computed at 0.25 and 0.775 sec; the window lengths are 250 milliseconds, and the dotted lines span that range in which the mean and standard deviation are computed. The higher of the two dotted lines represents the sum of the mean and some constant (as specified by the user) times the standard deviation of the region, while the lower line represents the mean less the same constant times the standard deviation. In this case, the coefficient of the standard deviation is very large, almost large enough that the two minimums barely get classified as glides.

A combination of absolute thresholds was combined with adaptive thresholds to determine the location of glides. Since the union of candidate regions was considered, the absolute thresholds chosen were conservative values. For example, needing to have F_2 larger than 2000 Hz or less than 800 Hz in order for a certain region to even possibly be classified as a glide region is more than reasonable because glides have much more extreme values. See Table 2.1 for glide formant frequencies. The locally adaptive window length was set to 250 milliseconds because measurements were made of glide transitions into vowels, and the durations seemed to vary between 75 and 175 milliseconds. Consequently, with the window centered on the glide landmark, it was most desirable to have the window contain the entire time frame up to the center of the vowel. 125 milliseconds (to one side of the landmark, 250 milliseconds total) appeared to be a reasonable average of the possible range of glide to vowel transition times. For F_1, F_2, and RMS amplitude, measurements were made from the VCV database of these parameters at the hand-labeled glide landmarks, and compared with the local mean and standard deviations. From these measurements, it was determined that most (over 95%) of the glide parameters fell outside of the following ranges: for F_1, $\mu_{F_1} \pm 0.4\sigma_{F_1}$; for F_2, $\mu_{F_1} \pm 0.7\sigma_{F_1}$; and for RMS amplitude, $\mu_{RMS} \pm 0.3\sigma_{RMS}$, where μ is the average and σ is the standard deviation of the data inside the 250 millisecond window.

So, for a given speech waveform, the 250 millisecond window was centered at the first data point (the first and last 125 milliseconds were mirrored, so that the first and last windows did not cover undefined regions; this did not affect the analysis because the beginnings and ends of waveforms usually contained silence), its F_1, F_2, and RMS amplitude individually compared with the local threshold, with the point being

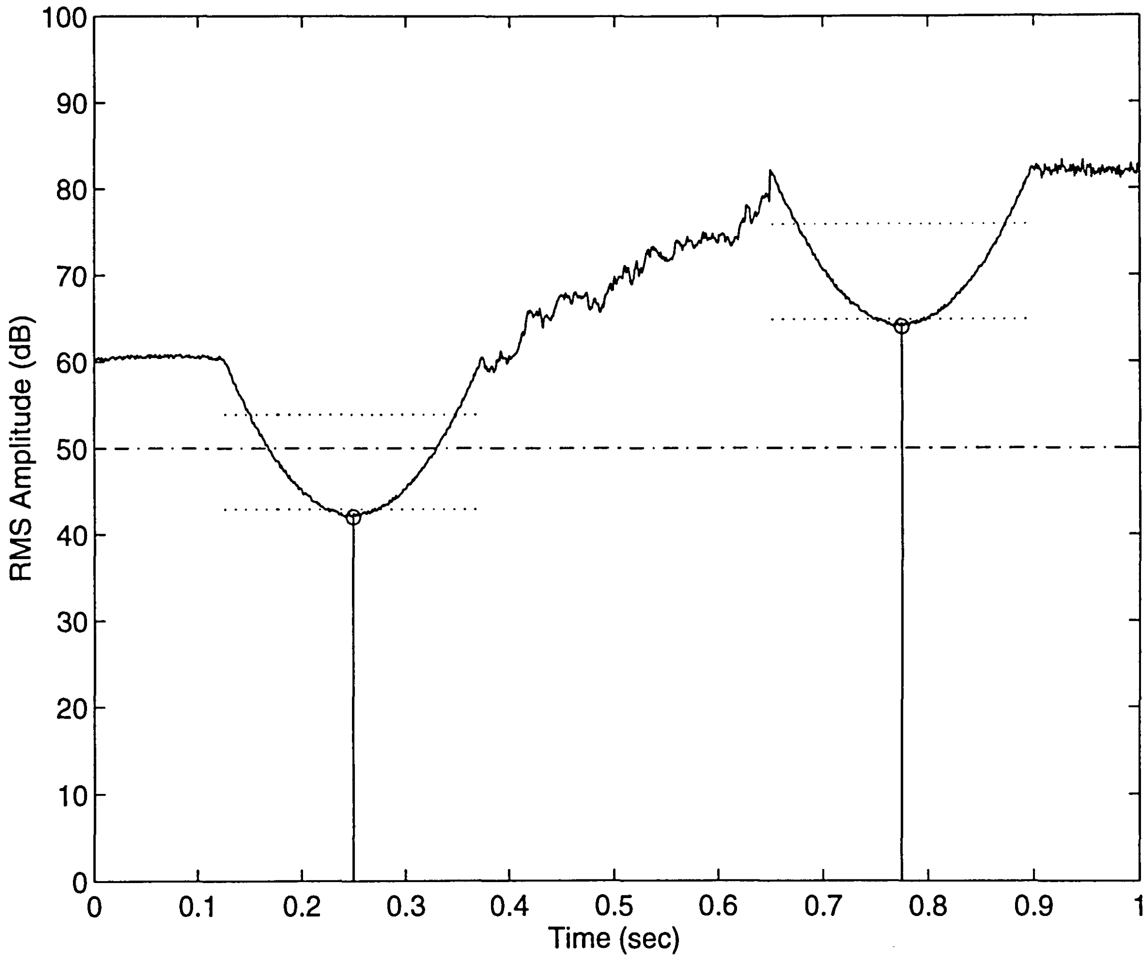

Figure 4-1: Example of how the adaptive thresholding works. Data generated is meant to represent a sample RMS amplitude plot over time. The plot itself is not an actual utterance.

Parameter	Value
F_1 k (coeff)	0.4
F_2 k (coeff)	0.7
RMS k (coeff)	0.3
F_1	less than 400 Hz
F_2	less than 800 Hz or more than 2000 Hz
Window Length	250 msec
Sample Updates	every 2.5 msec

Table 4.1: Parameter values chosen for glide detection analysis

marked as a potential glide if all three parameters (F_1, F_2, and RMS amplitude) were outside the threshold ranges. Then, the window was moved down 2.5 millisecond, and the process was repeated. It should be noted that no prior labels were used in this analysis; so, this algorithm, if successful, would have been self-sufficient. At the end, certain regions of the speech waveform were marked as possible glide locations. Next, the absolute measures were implemented. Absolute thresholds of F_1 less than 400 Hz, F_2 less than 800 Hz or greater than 2000 Hz were run on the waveforms to mark another set of possible glide points. The two sets of thresholds (local and global) were then combined together, and points in time where both methods found a possibility of a glide location were marked as glide locations. Since glides have slow transitions, only regions which had 10 milliseconds (five consecutive points) marked as glides were considered as legitimate glide regions, and the landmark was located at the center of this set of points. Regions with four or less consecutive points marked as glides were deemed points found from noise or formant tracking errors (more on this on tracking errors will be discussed below). Table 4.1 summarizes the thresholds used on each parameter.

4.1.3 Results

This algorithm was run on the sentences database, resulting in an accuracy of 71% for properly locating glides. Further, approximately 1.1 improperly labeled glide points were made in each of the sentences, on the average. Such a result made the author decide that other approaches could be found to obtain better results. Errors

Time (sec)	Glide	Error?
Correct Sentence		
0.24	/w/ (in *we*)	no
0.48	/w/ (in *were*)	no
0.75	/w/ (in *away*)	no
1.16	/j/ (in *year*)	no
Sentence with Errors		
0.31	/j/ (in *you*)	yes
0.48	/w/ (in *went*)	no
1.20	/w/ (in *wire*)	no
2.09	/w/ (in *Gwen*)	yes
2.43	/j/ (in *yacht*)	yes

Table 4.2: Legend for glide labels in two sentences shown in Figure 4-2 and Figure 4-3.

in formant tracking were the primary cause of errors in this detection process.

Figure 4-2 shows an example where the formant tracking worked fine, or reasonably well, as was often the case. The three lines represent F_1, F_2, and F_3, with F_1 being the line with the lowest frequency, F_2 being the line in the middle of the two other lines, and F_3 being the line on top. The lollipops represent glide locations hand-labeled. In cases such as the one shown in Figure 4-3 (again, the lines represent the three formants, as in Figure 4-2, and the lollipops represent glide locations), the formant tracking failed in the region of the glides. This error was a result of the amplitude of the signal being so small that the formant tracking routine of Entropic could not track it. Table 4.2 provides information regarding the particular glide labeled by the lollipops in the figures. In the second sentence, F_1 has just barely found its correct value while F_2 is still searching for the actual F_2 at the location of the /j/ in *your* (time 0.31 seconds). Likewise, at 2.09 and 2.43 seconds, F_1 is unreliable, and F_2 for the former is wavering between F_2 and F_3 tracks.

4.1.4 Drawbacks

In fact, if these tracking errors were removed from the list of errors (leading mostly to improper labels of glides), the average of improper labels would be less than 0.9 per sentence. The errors in formant tracking occurred in regions of low energy; such as in a glide, and more so in stops and voiceless fricatives. In these cases, the formant tracker

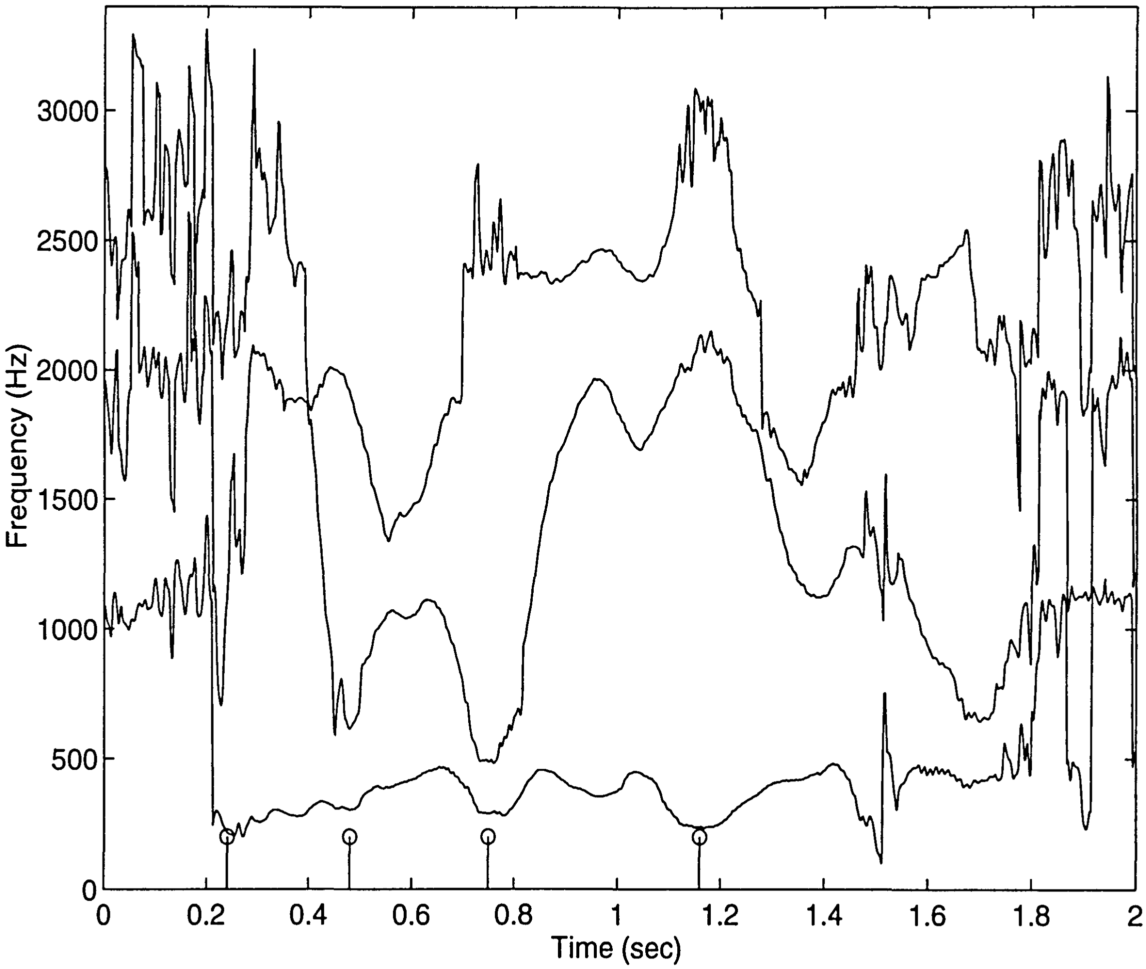

Figure 4-2: Example where automatic formant tracking from Entropic tools worked in the region of the glides. Utterance was *We were away a year ago.* Table 4.2 provides a legend for the lollipops shown.

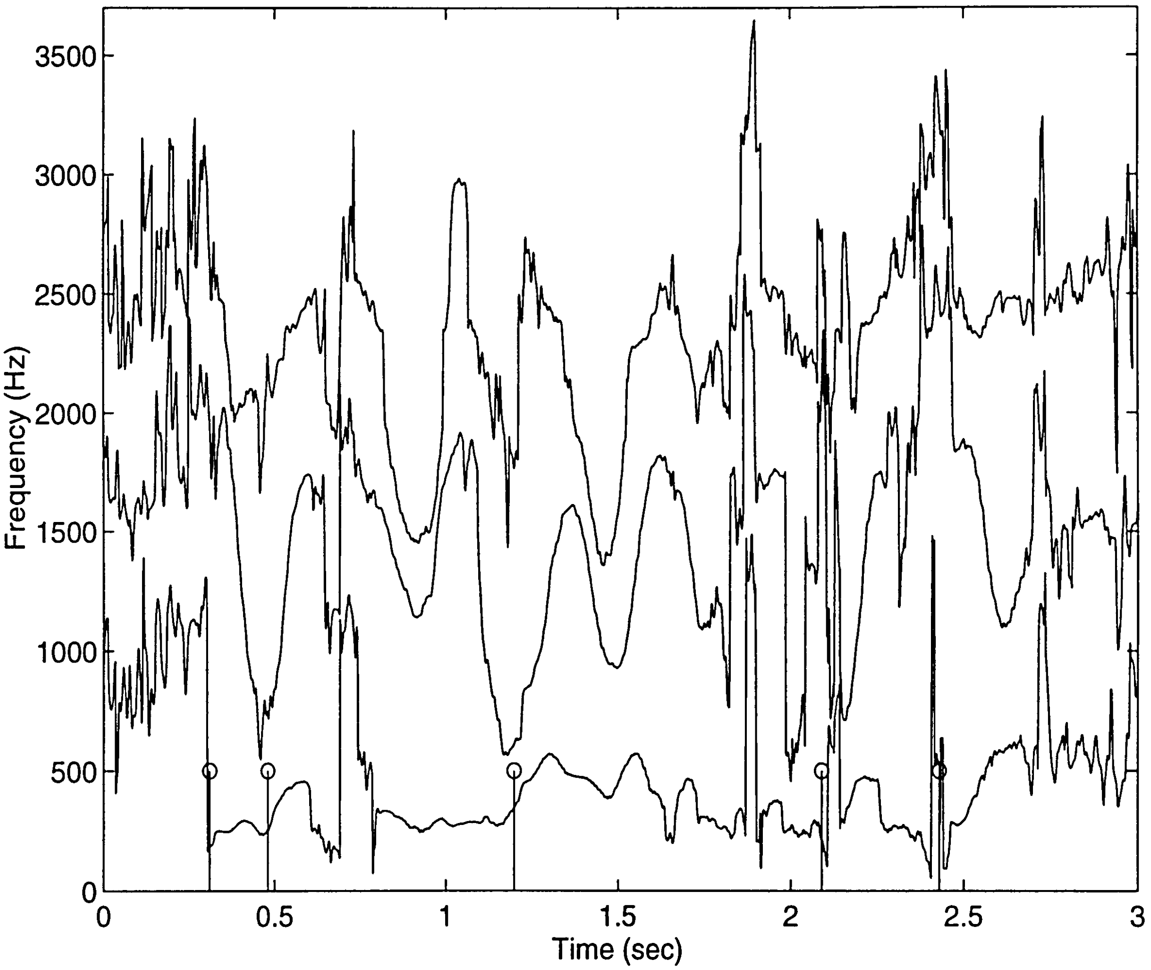

Figure 4-3: Example where automatic formant tracking from Entropic tools failed in the region of the glides. Utterance was *You went to rewire items in Gwen's yacht.* Table 4.2 rovides a legend for the lollipops shown.

would sometimes abruptly jump to the next higher formant if the lower formant track was lost. And in regions between closure and release, where discussing formants is irrelevant, sometimes the tracker would move along so that its values would satisfy the constraints for a glide. This would result in a false find of a glide because of its rapid deviation from the local average of the waveform.

Thus, errors in formant tracking played a major role in why this technique was not pursued further. However, this does not mean that 100% detection and no false glide labels would be obtained if a perfect formant tracker were used, since the thresholding on the RMS amplitude data on occasion failed to locate glides.

4.2 Modeling the Transition as a Sinusoid

A characteristic of a glide is that it has an initial brief steady-state for F_1 and RMS amplitude, which moves up gradually, and then reaches another steady-state when it approaches the following vowel. This movement appears similar to the movement of a sinusoid from its minimum to its maximum. Thus, a method to find a best-fit sinusoid on F_1 and RMS amplitude on the transition from the RMS amplitude minima to the center of the following vowel was made, with mean-square error being the error criterion to be minimized.

4.2.1 Obtaining Measurements

The data desired for this analysis was F_1 and RMS amplitude. F_1 was obtained using the formant tracker from Entropic Signal Processing Systems, while RMS amplitude was found using a 49 millisecond rectangular window. The center of the vowel was labeled as the point of maximum RMS amplitude and used as one edge for the region of analysis, and the minimum of RMS amplitude in the consonant was labeled and used as the other edge of the region. However, having these points as the endpoints did not force the sinusoid to have zero slope at both ends (only at the end of the vowel). The frequency of the sinusoid was left as a parameter, also; so, the best fit line was sought with only the single constraint that the right endpoint had to be a maximum. Other methods of free parameters could have been chosen. The ones discussed were the ones chosen for this analysis. To reduce the effects of noise, a ten millisecond moving averager was performed on the data before the best fit was made.

4.2.2 Parameters Used in Sinusoidal Best Fit

The maximum of the sinusoid was fixed on the right end where the middle of the vowel was. The free parameters allowed were the DC offset, the coefficient on the sinusoid, and the frequency of the sinusoid. The frequency, however, was restricted to the range of 2.86 Hz to 6.67 Hz because this region maps to a half-sinusoid period range of 75 milliseconds to 175 milliseconds, the range of possible glide durations as measured and discussed in the previous section (4.1.2).

4.2.3 Determining the Best Fit

Following is the definition of mean-square error (MSE) for the problem at hand.

$$MSE = \sum_{i=0}^{n}(A - Bcos(\omega t_i) - x_i)^2 \tag{4.1}$$

t_i are the times the data was sampled, and x_i are the sampled data values. Now, the process was to find A, B, and ω to minimize the MSE. These constants were solved for my taking derivatives with respect to each of the variables and setting them equal to zero.

$$\frac{d}{dA}MSE = 2(nA - B\sum_{i=0}^{n} cos(\omega t_i) - \sum_{i=0}^{n} x_i) = 0 \tag{4.2}$$

$$\frac{d}{dB}MSE = 2(A\sum_{i=0}^{n} cos(\omega t_i) - B\sum_{i=0}^{n} cos^2(\omega t_i) - \sum_{i=0}^{n}(x_i cos(\omega t_i))) = 0 \tag{4.3}$$

$$\frac{d}{d\omega}MSE = 2B(A\sum_{i=0}^{n} t_i sin(\omega t_i) - B\sum_{i=0}^{n} \frac{t_i}{2} sin(2\omega t_i) - \sum_{i=0}^{n} x_i t_i sin(\omega t_i)) = 0 \tag{4.4}$$

The system of equations is non-linear, and a closed form solution would be difficult to find. However, the closed form solution for A and B, given a constant ω, could be found easily. As a result, A and B were found for constant ω, and in the m-file written, different values of ω were chosen, with the minimum MSE being found by finding which value of ω gave such a result. With this restriction, the closed form solutions for A and B were found to be

$$B = \frac{n\sum_{i=0}^{n} x_i cos(\omega t_i) - \sum_{i=0}^{n} x_i \sum_{i=0}^{n} cos(\omega t_i)}{(\sum_{i=0}^{n} cos(\omega t_i))^2 - n\sum_{i=0}^{n} cos^2(\omega t_i)} \tag{4.5}$$

$$A = \frac{1}{n}(B\sum_{i=0}^{n} cos(\omega t_i) + \sum_{i=0}^{n} x_i) \tag{4.6}$$

The code used to solve for the best sine fit may be found in Appendix A.2.

4.2.4 Results

Examples of curve fits may be seen in Figure 4-4, 4-5, and 4-6. Figure 4-4 shows that the RMS transition for a glide (from a /w/ to a /aa/) fits a sine curve well. Figure 4-5 shows the case where the F_1 of the glide leads to a reasonable fit (it is the transition from a /j/ into the vowel /aa/); so, it can been seen that a best sine fit on the formants works well, too. Figure 4-6 shows the poor sine fit for a transition of a non-glide into the vowel (a /dj/ into an /aa/). As can be seen, the fits for the glides into the vowel appear to be quite sinusoidal, while for the example provided, the non-glide into the vowel does not fit a sine.

4.2.5 Drawbacks to This Method

The results shown make this idea somewhat appealing, but the drawback is that many non-glide consonants have fits which are sinusoidal. This is particularly true for liquids and unvoiced fricatives. An example of this may be found in Figure 4-7, which shows that the sinusoid model fits the particular /r/ transition into the vowel /eh/ well. Furthermore, the lack of a perfect formant tracker prevents assurance that F_1 best fits are made on accurate formant tracks. Consequently, this method was also abandoned for the time being. With a good formant tracker, plus a more thorough analysis of varying the parameters, one could possibly obtain better results using this method.

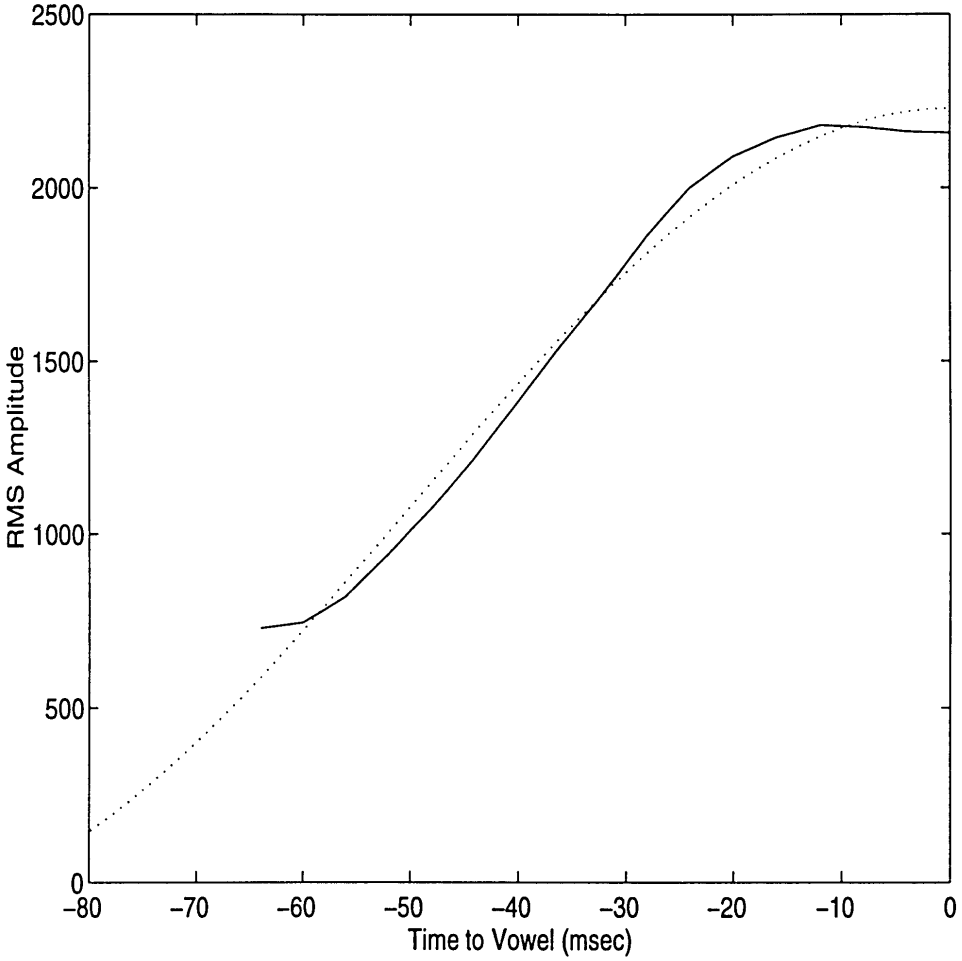

Figure 4-4: Best fit sine (dotted line) along with the actual RMS amplitude (solid line) for the glide /y/ going into the vowel /aa/

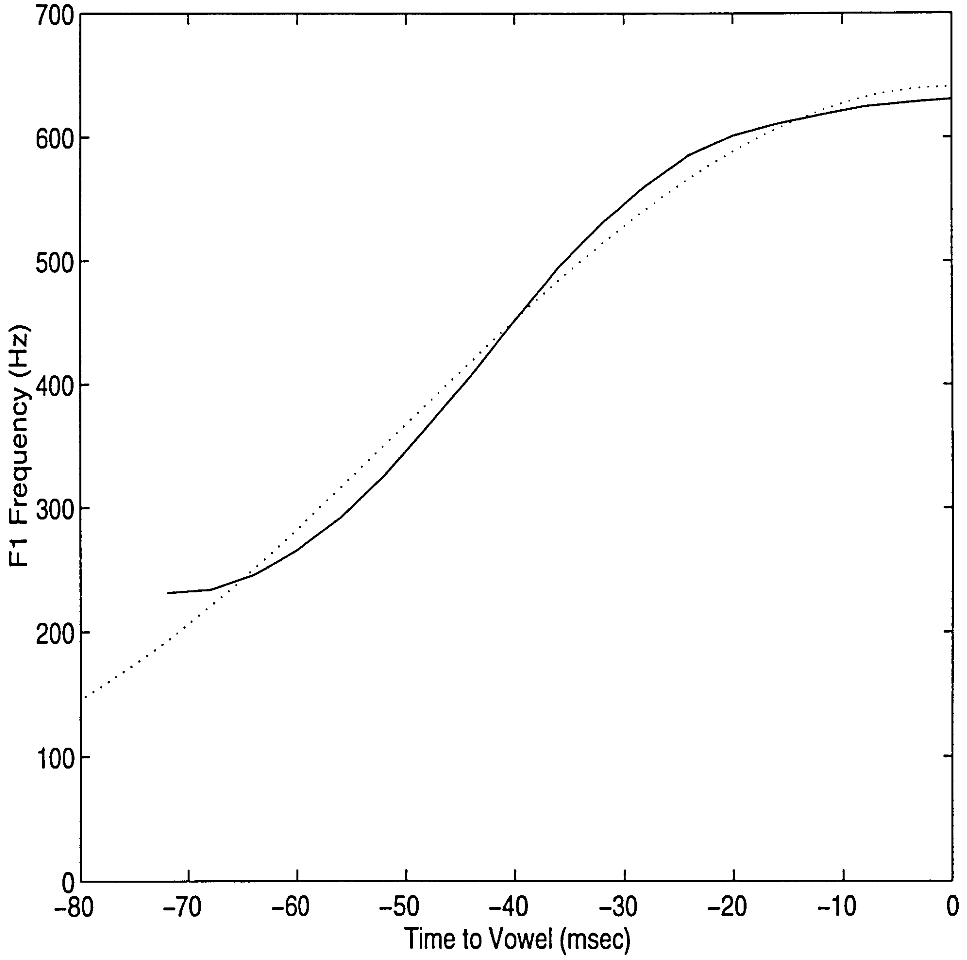

Figure 4-5: Best fit sine (dotted line) along with the actual F_1 curve (solid line) for the glide /w/ going into the vowel /aa/

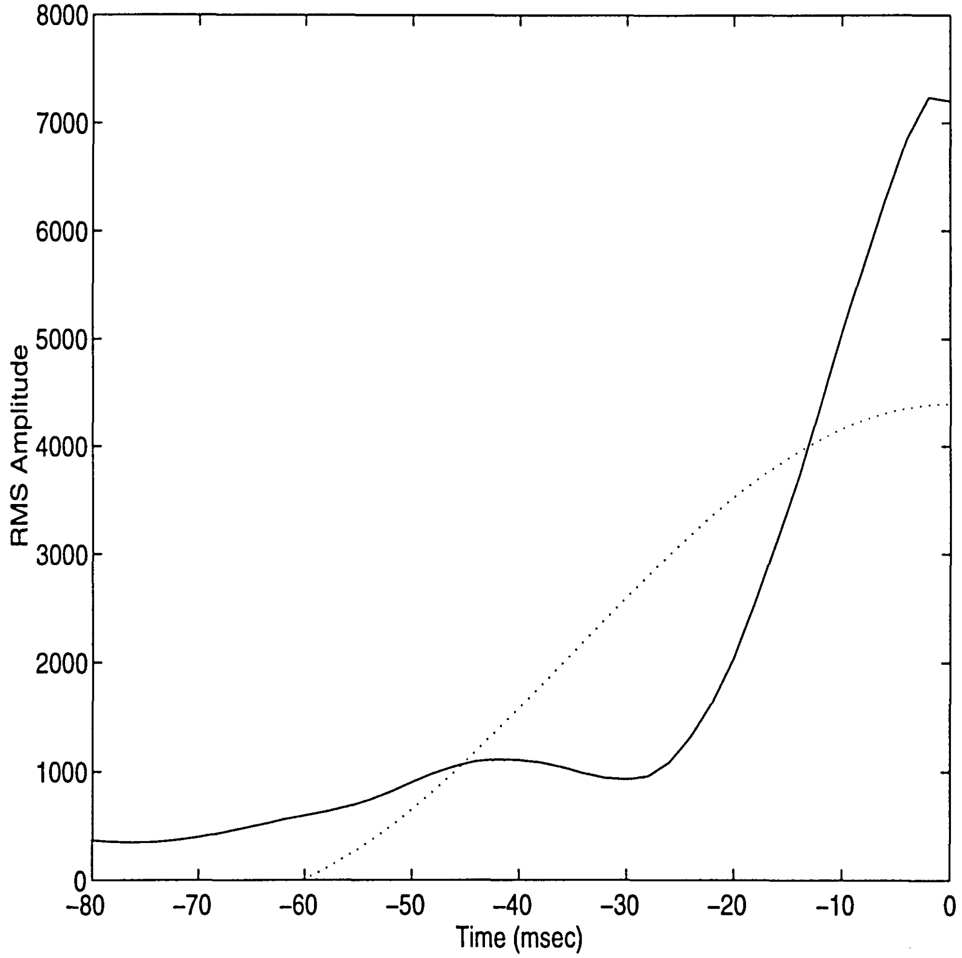

Figure 4-6: Best fit sine (dotted line) along with the actual RMS amplitude (solid line) for /dj/ into the vowel /aa/. As expected, this fit is not good because the previous phoneme is not a glide.

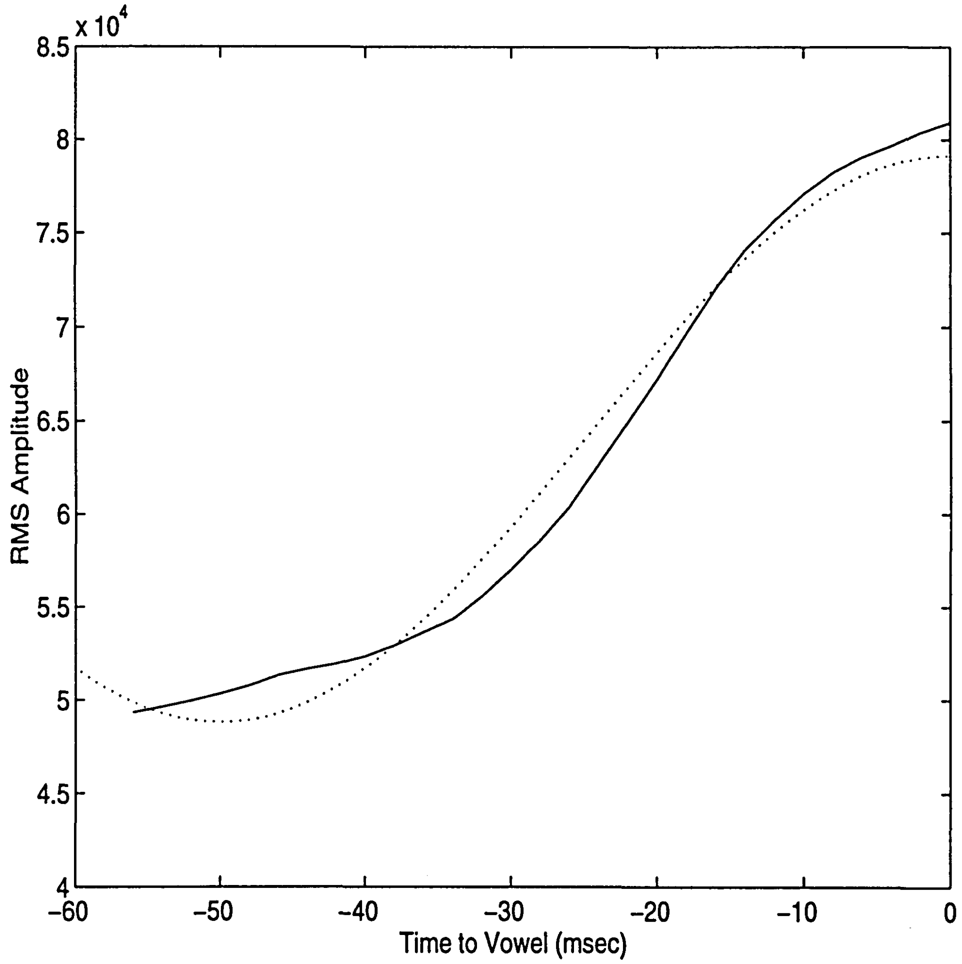

Figure 4-7: Best fit sine (dotted line) along with the actual RMS amplitude (solid line) for /r/ into the vowel /eh/. The sinusoid fits well even though it is not a glide.

Chapter 5

Development of Recognition System

5.1 VCV Data

The development of the final algorithm began with an analysis of the VCV database as a training set to extract particular acoustic properties. After the data were collected, information was organized according to different vowel contexts (aa, ah, eh, iy, and uw). The maximum rate of change of the RMS amplitude along with the maximum dB range of RMS amplitude were measured for the VCVs. Furthermore, the same two measurements were made for F_1. In all of these measurements, thirty millisecond shifts forward from the landmark were made before analysis. This number was chosen because the landmark detector used to locate release points of obstruent consonant landmarks was accurate to within ± 30 milliseconds [Liu, 1995 (Ch. 3.3)]. Figure 5-1 shows a sample transition of both a glide and a non-glide. As one can see, the sharp transition occurs well within the first thirty milliseconds, while most of the smoothness of the glide transition can still be seen after thirty milliseconds.

Plots of amplitude and frequency range versus maximum rates of change were made for RMS and F_1, respectively, and the means and covariance matrices were computed separately for glides and non-glides. This information was then used as the probability distribution models for glides and non-glides for each particular con-

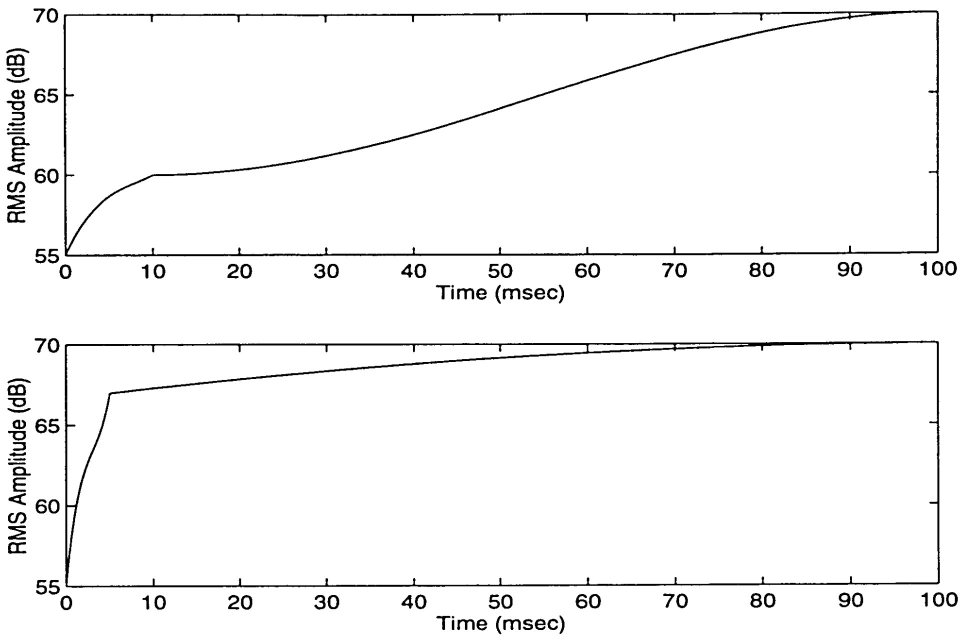

Figure 5-1: Example of RMS transitions for (top) glide between consonant release and vowel (abrupt consonant-glide-vowel utterance) and for (bottom) no glide from release to vowel (abrupt consonant-vowel utterance). The abrupt landmark release is at time t=0; notice how the transition with the glide is much smoother than the one without the glide.

text. Using these models, a hypothesis test was used to determine whether the given properties were more like that of a glide, or a non-glide.

Figure 5-2 shows a scatterplot of RMS amplitude range versus the maximum rate of change of RMS amplitude for all of the VCV data. The dots represent non-glide tokens, while the G's represent glide tokens.

Some overlap can be seen between the glides and non-glides. This overlap could be better separated by looking at the context more carefully, both of the following vowel and the particular non-glide consonant. Following will be an analysis of how different types of consonants compare to the glides. Afterwards, a discussion of the effects of different following vowels will be given.

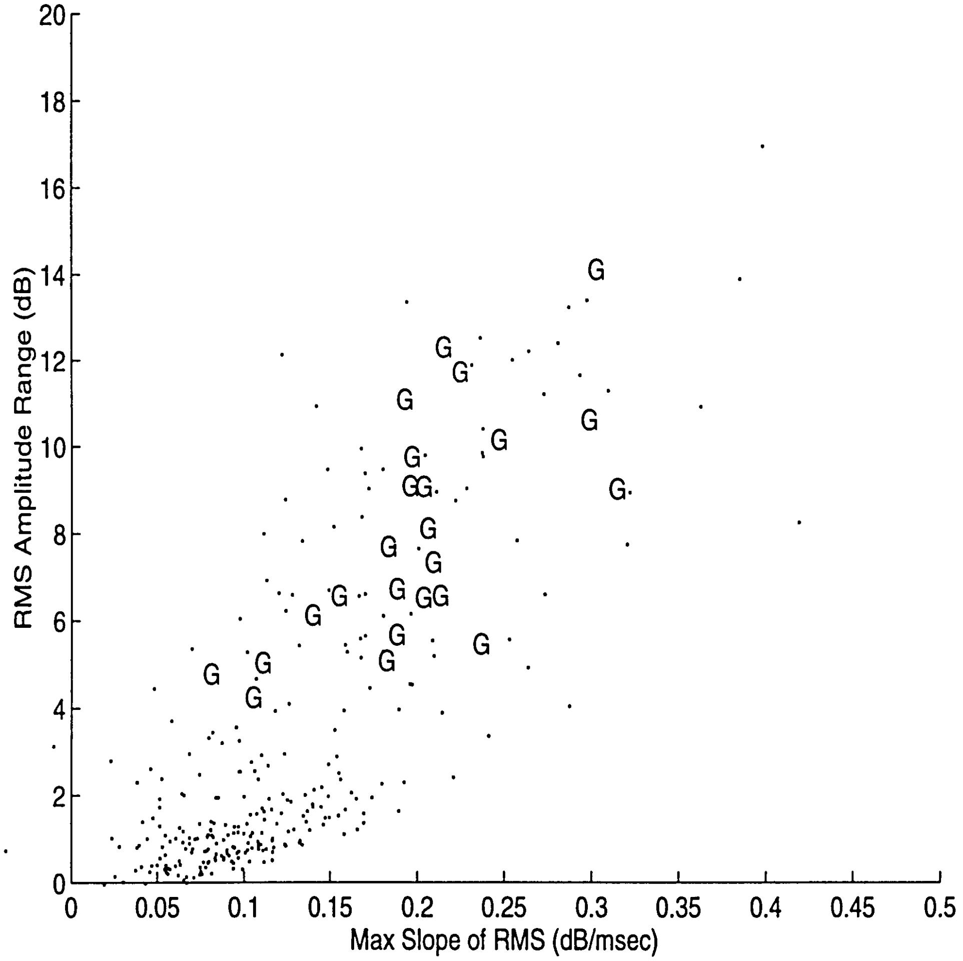

Figure 5-2: Scatterplot of VCV data; dB versus dB/msec. G = Glide, . = Non-Glide.

5.2 Comparing Consonants

A comparison of each class on consonants with the glides was made from the VCV database. Following are the results for RMS amplitude.

5.2.1 Affricates

The affricates /dj/ and /ch/ in English are a mix between stops and fricatives. They have a closure and release, but the release is filled with frication; hence the name. The production of an affricate is much different than that for a glide, so one would expect its characteristcs to be much different. As expected, such is the case, with affricates generally having a much lower range and rate of change of RMS amplitude. Figure 5-3 shows this result graphically.

5.2.2 Voiced Stops

The voiced stops /b/, /d/, and /g/ all have a closure and an abrupt release, characteristic of all stops. After thiry milliseconds from the release, the RMS amplitude is just about at the steady state value. As a result, the dB range and the maximum RMS amplitude change are much smaller than those found in glides. Figure 5-4 illustrates this. As was the case for affricates, the voiced stops are also easy to distinguish as non-glides.

5.2.3 Unvoiced Stops

The unvoiced stops /p/, /t/, and /k/ are similar to the voiced stops, except for voicing. This difference amounts to a period of voiceless aspiration out of the release. Again, after thirty milliseconds from the onset of voicing, the energy in a voiced stop is practically at its final value. Thus, the RMS amplitude range is small, along with a small rate of change of RMS amplitude at the release. As Figure 5-5 indicates, separation of glides with unvoiced stops is clear.

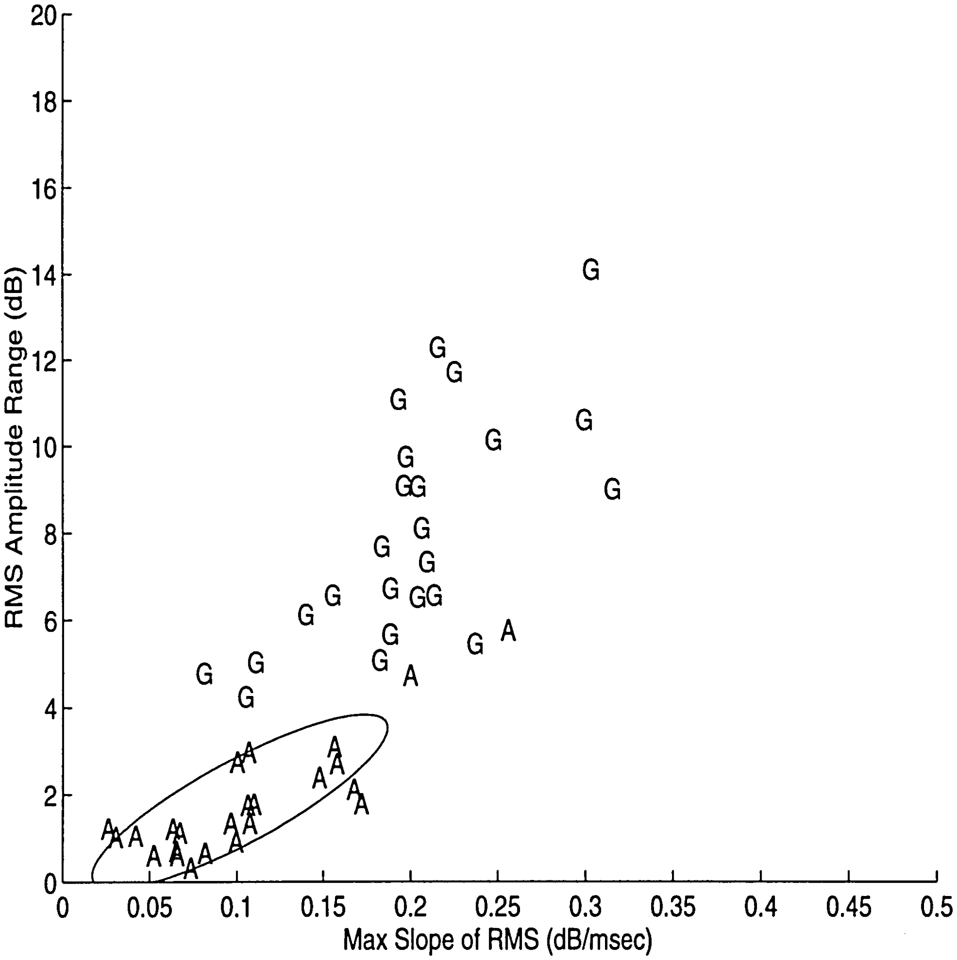

Figure 5-3: Scatterplot of Glides and Affricates; dB versus dB/msec. G = Glide, A = Affricate (1.5 standard deviation curve for the affricates).

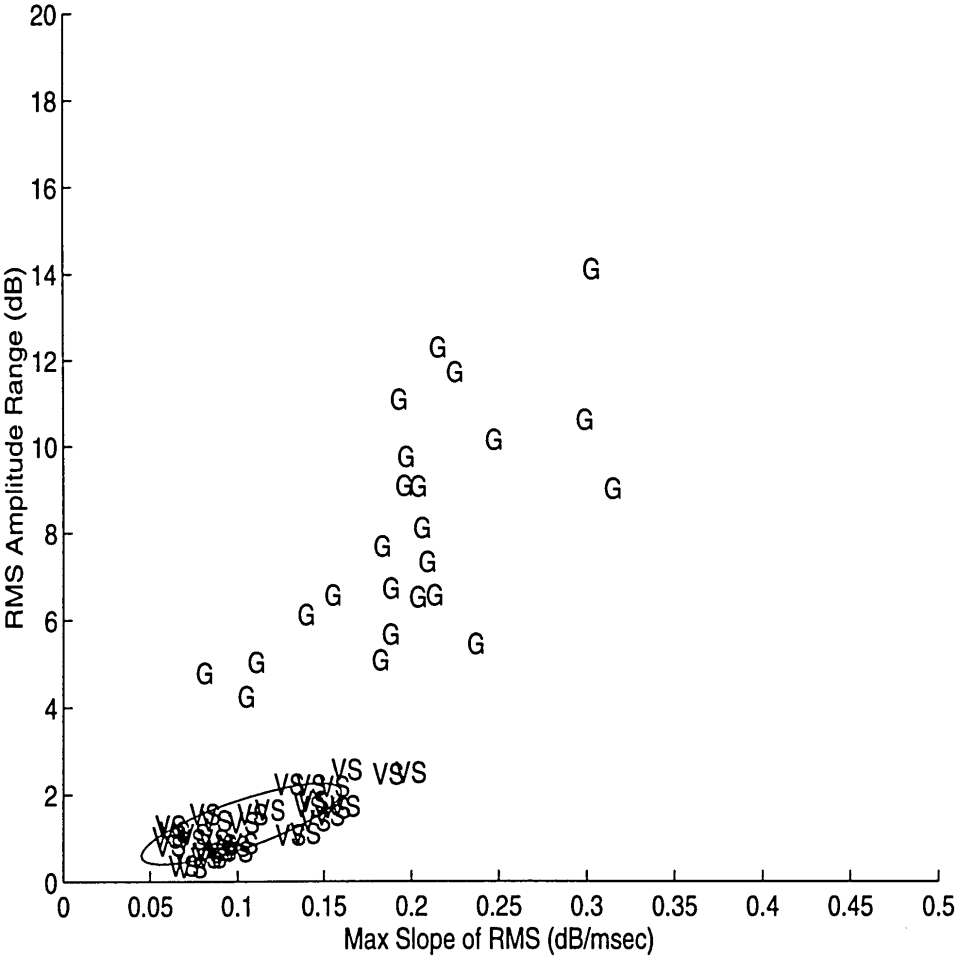

Figure 5-4: Scatterplot of Glides and Voiced Stops; dB versus dB/msec. G = Glide, VS = Voiced Stop (1.5 standard deviation curve for the voiced stops).

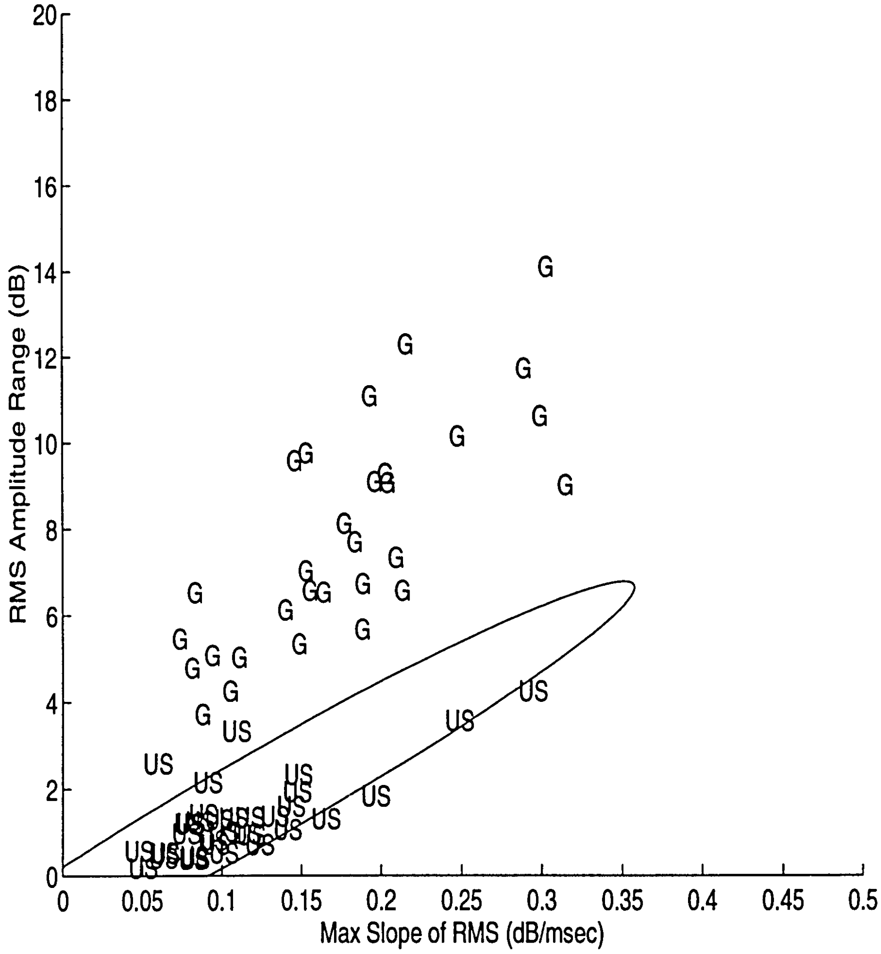

Figure 5-5: Scatterplot of Glides and Unvoiced Stops; dB versus dB/msec. G = Glide, US = Unvoiced Stop (1.5 standard deviation curve for the unvoiced stops).

5.2.4 Voiced Fricatives

The voiced fricatives /dh/, /v/, /z/, and /zh/ have characteristics similar to those of glides. Both have a constriction within the vocal tract, and both are voiced. The only difference is that fricatives have noisy spectra, whereas glides do not. Unfortunately, sometimes this difference does not lead to significant differences in maximum slope and dB range characteristics, as seen in Figure 5-6. In addition, sometimes the landmarks in voiced fricatives are not properly located, leading to erroneous data being computed. The voiced fricatives with correctly labeled landmarks lie in a region close to the bottom left of the plot, while the remaining points are due to errors. As a result, separating voiced fricatives from glides is more difficult than any of the other types of consonants listed above. The errors made were primarily confusions of /dh/ and /v/, since these two fricatives most closely resemble glides when no abruptness is produced while uttering these consonants. It will be seen in the following chapter of results on the sentences database that voiced fricatives are one class of consonants in which the recognition algorithm occasionally classifies as a glide.

5.2.5 Unvoiced Fricatives

The unvoiced fricatives /f/, /s/, /sh/, /th/, on the other hand, are not as difficult to separate from the glides. Although these consonants are made with a constriction, the lack of voicing makes it easier to distinguish from glides. Figure 5-7 shows the distribution of unvoiced fricatives compared to glides.

5.2.6 Nasals

The nasals /m/, /n/, and /ng/ belong to a larger class of sonorants to which glides belong, also. Sonorants are sounds produced without a build-up of pressure in the vocal tract. As a result, large changes in RMS amplitude are generally not seen in any of the sonorants, although changes in amplitude in higher frequency ranges may be more abrupt. As a result, nasals and glides have very similar features. Figure 5-8 compares the two types of sonorants. It can be observed that differences between the

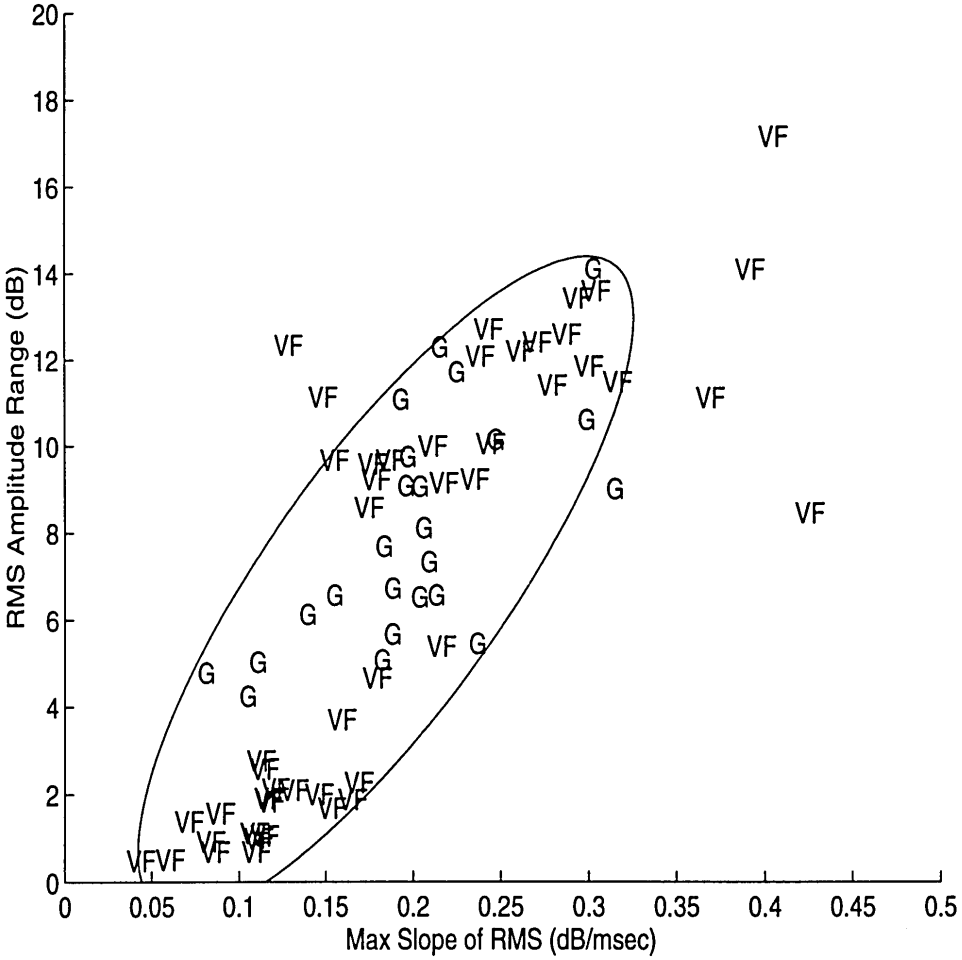

Figure 5-6: Scatterplot of Glides and Voiced Fricatives; dB versus dB/msec. G = Glide, VF = Voiced Fricative (1.5 standard deviation curve for the voiced fricative).

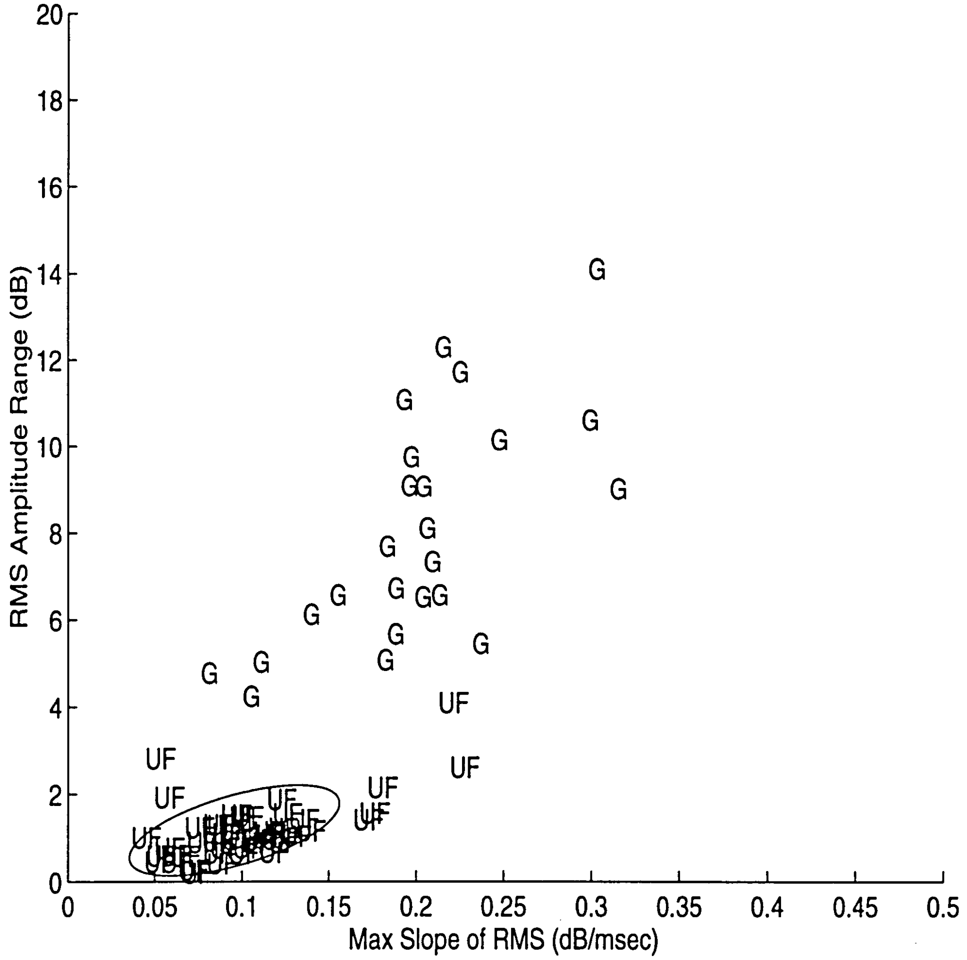

Figure 5-7: Scatterplot of Glides and Unvoiced Fricatives; dB versus dB/msec. G = Glide, UF = Unvoiced Fricative (1.5 standard deviation curve for the unvoiced fricatives).

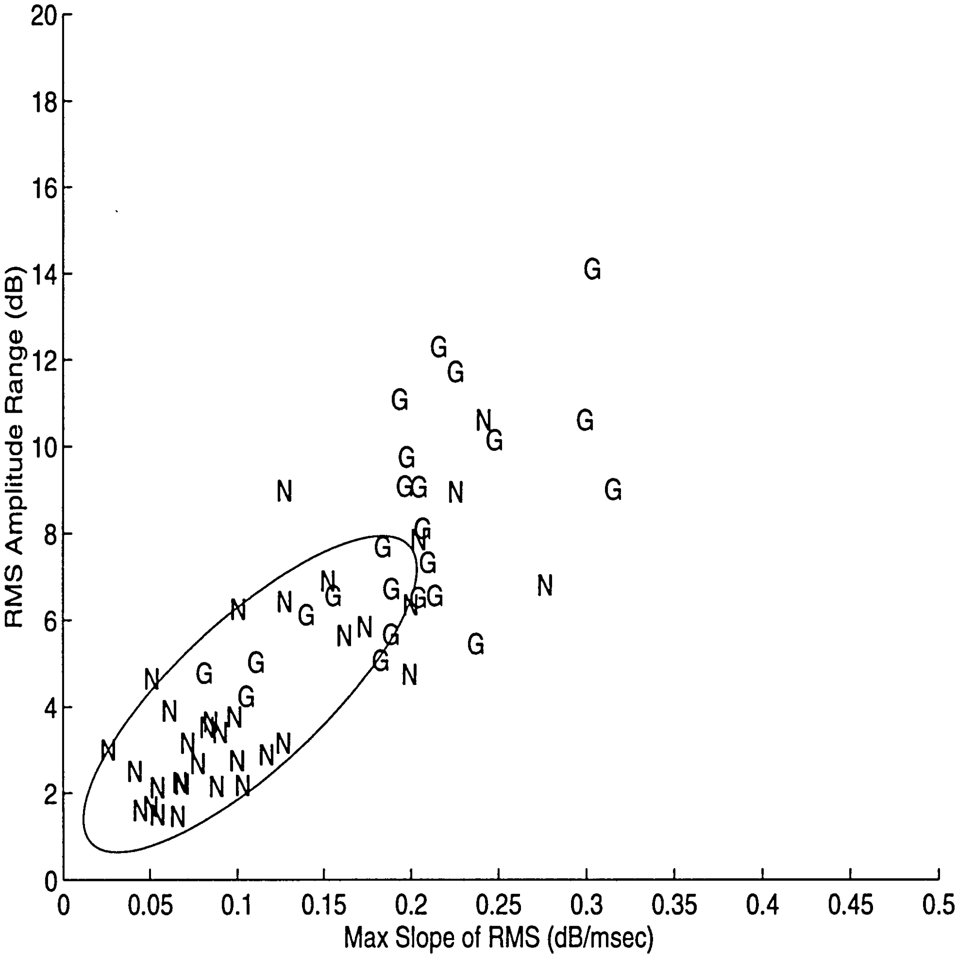

Figure 5-8: Scatterplot of Glides and Nasals; dB versus dB/msec. G = Glide, N = Nasals (1.5 standard deviation curve for the nasals).

nasals and glides are difficult to detect using the two characteristics (max slope of RMS amplitude, and RMS amplitude range). From the graph, one may observe that nasals generally have a smaller amplitude range, but not significantly smaller. The only major difference that aids in locating nasals is an abrupt change in low frequency energy first due to the closure of the major articulator (lips, tongue blade, or tongue body) and then the opening of the velopharyngeal port.

5.2.7 Liquids

The liquids /l/ and /r/ pose a problem similar to that of nasals because they are also sonorant consonants. In fact, liquids are even more similar than nasals because

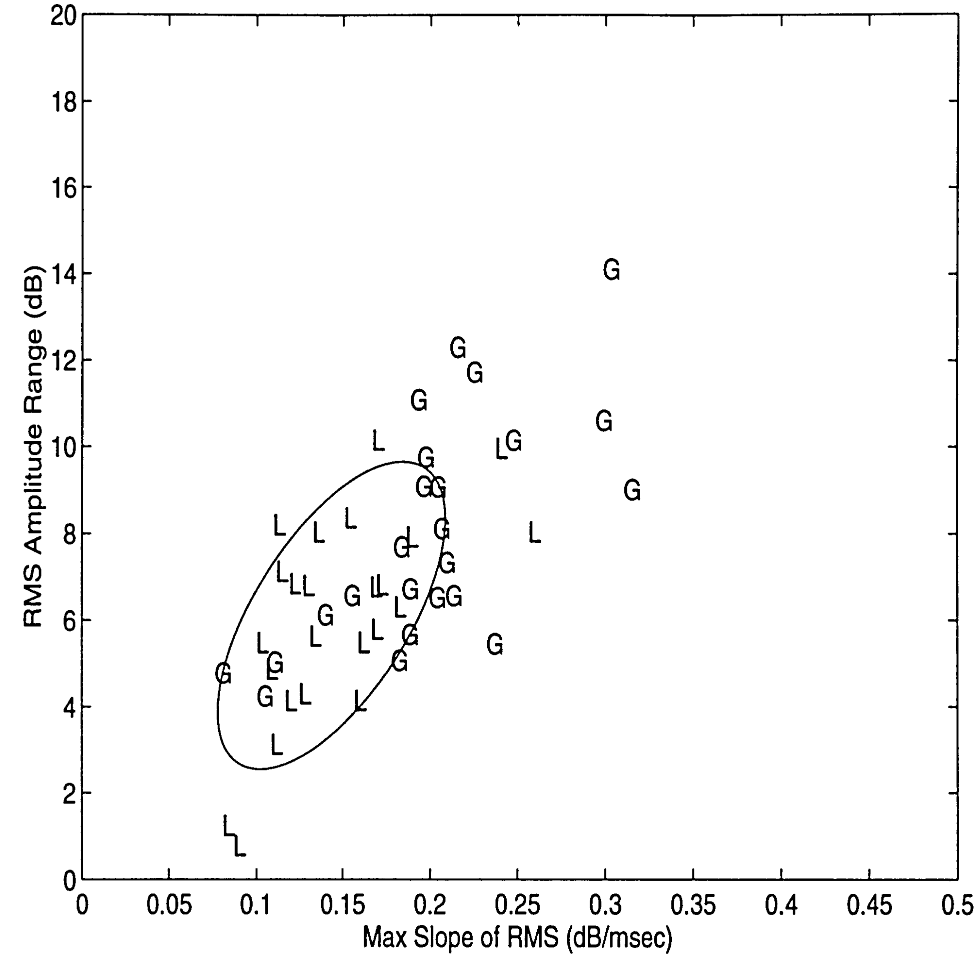

Figure 5-9: Scatterplot of Glides and Liquids; dB versus dB/msec. G = Glide, L = Liquids (1.5 standard deviation curve for the liquids).

the energy into and out of a liquid is often smooth. In fact, observation of Figure 5-9, which shows the plot of liquids and glides for the VCVs, shows that almost no separation can be found between these two classes of sonorants.

5.3 Effects of Vowel Context

The particular vowel following the consonant also played a role in the measurements. This effect was expected for the following two reasons. First, different vowels have different amplitudes [Peterson and Barney, 1952]. Table 5.1 shows the amplitude measurements of F_1 made by Peterson and Barney. Furthermore, since the energy of a glide or a vowel is dominated by the low frequency energy, F_1 amplitude should

Vowel	relative F_1 amplitude (dB)
aa	-1
ah	-1
eh	-2
uw	-3
iy	-4

Table 5.1: Relative F_1 amplitudes of vowels as reported by Peterson and Barney [1952].

Vowel	Position (High/Low)	F_1 (Hz) adult male	F_1 (Hz) adult female
aa	Low	730	850
ah	Low	640	760
eh	Neither	530	610
uw	High	300	370
iy	High	270	310

Table 5.2: F_1 frequencies of vowels as reported by Peterson and Barney [1952].

be a good measure of the overall energy of the signal. From Peterson and Barney's measurements, one can see that vowels such as /aa/ have two and three decibels more energy than /uw/ and /iy/, respectively.

Second, low vowels have high F_1, while high vowels have low F_1. As a result, the change in frequency from a glide to the following vowel can vary substantially based on the final steady-state value of the following vowel. Table 5.2 shows the F_1 values as measured by Peterson and Barney [1952].

Figure 5-10 and Figure 5-11 show the distributions of the different vowels in a plot of F_1 range versus maximum F_1 rate of change for /w/ and /j/, respectively. As one can observe, the /uw/ and /iy/ tokens are near the bottom of the plot, while those for /aa/, /ah/, and /eh/ are closer to the top. This can be attributed to the fact that F_1 does not have to move up as far from the glide into the high vowels. It should also be noted that even though the tokens for different vowels are separated, they all tend to lie along a straight line through the origin, as witnessed in the graphs by the high eccentricity (close to one) of the ellipse. This indicates that the points lie all along one direction. Consequently, one can conclude that each token has a similar ratio of

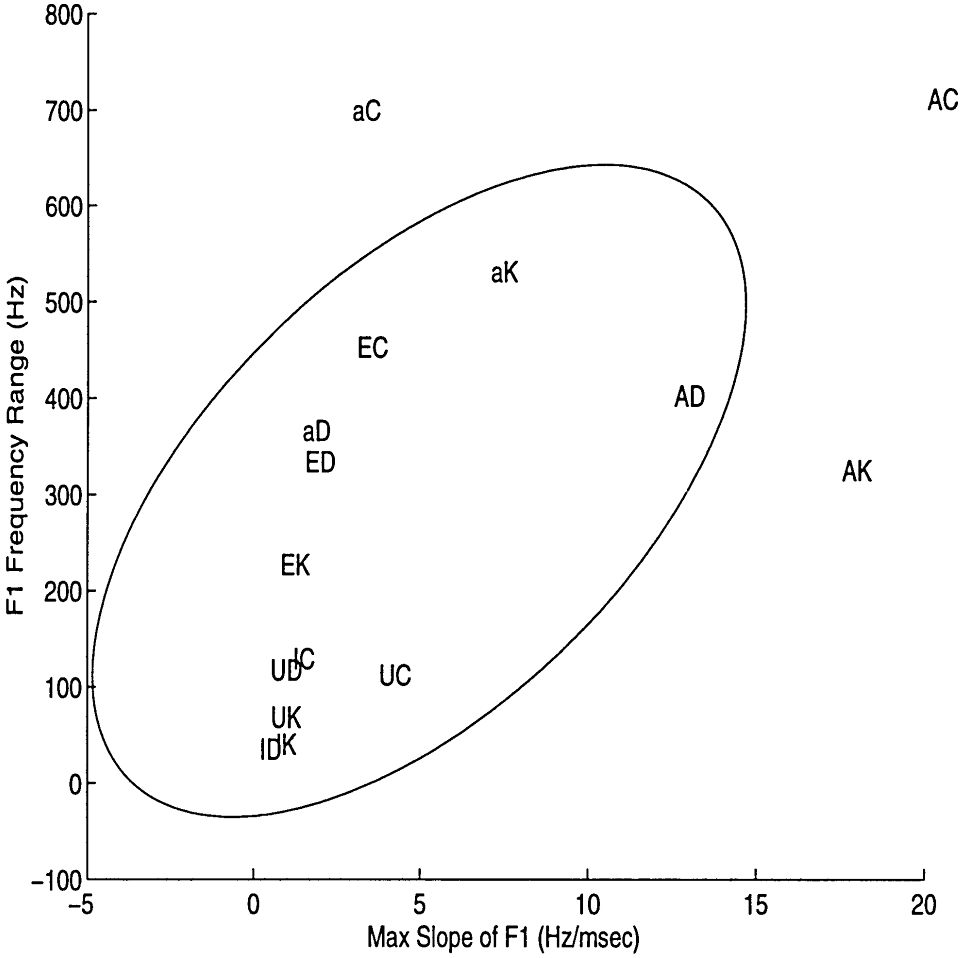

Figure 5-10: Plot of F_1 distributions for /w/s, along with 1.5 standard deviation curve. First letter in the labels is the following vowel (A=aa, a=ah, E=eh, I=iy, U=uw), while the second letter is the speaker (C=CB, D=DW, K=KS).

F_1 range to maximum rate of change. This fact illustrates that the time it takes to go from the glide into the following vowel is more or less constant. As a result, the rate of movement is adjusted to conform to the constant duration. This result is observed also for RMS amplitude data. A quantitative discussion of this information follows in the next section.

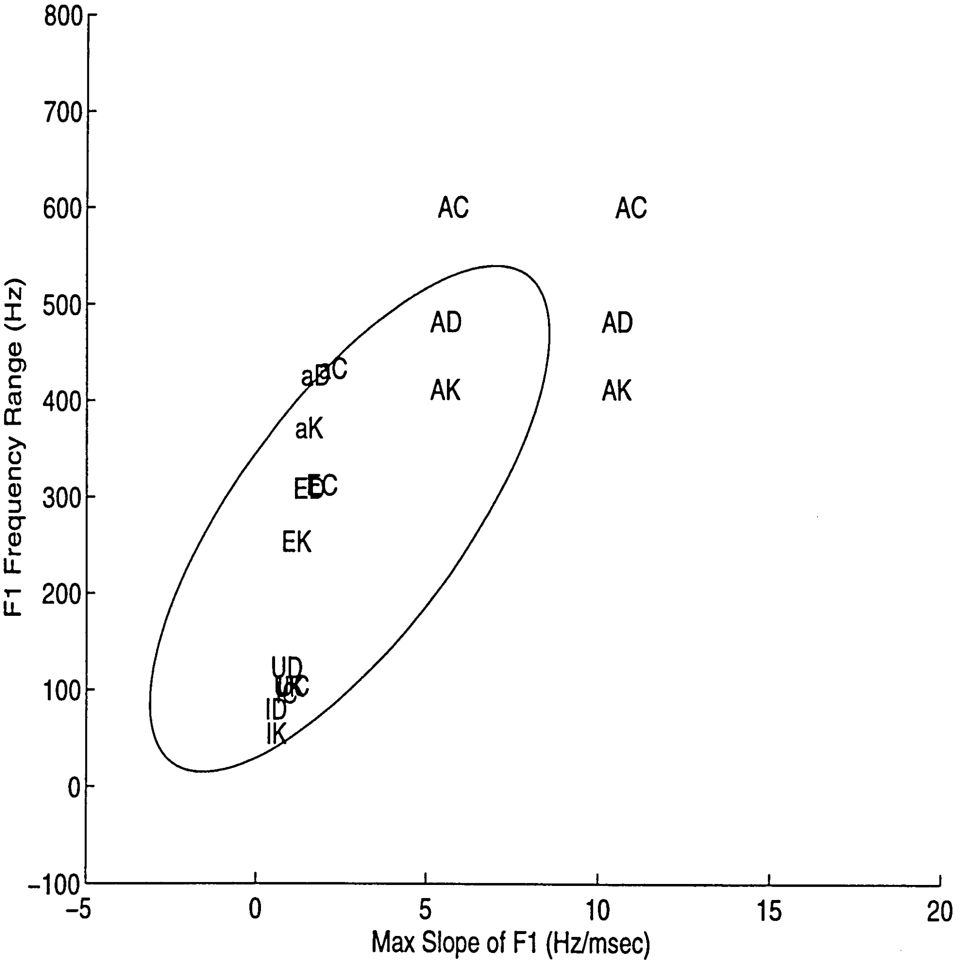

Figure 5-11: Plot of F_1 distributions for /j/s, along with 1.5 standard deviation curve. First letter in the labels is the following vowel (A=aa, a=ah, E=eh, I=iy, U=uw), while the second letter is the speaker (C=CB, D=DW, K=KS).

5.3.1 Determining Glide to Vowel Duration using a Sinusoidal Model

If one assumes a sinusoid model for the transition from the glide to the vowel, one can determine the duration from the glide to the following vowel on an RMS or F_1 range versus maximum rate of change plot by finding the slope of the line which the points more or less lie on. This can be seen by the following: if one wants to model the transition as a sinusoid, then one can write write the sinusoid as -D cos($\frac{2\pi t}{2T}$), where T is the time from glide to vowel, the glide begins at time t=0, and the maximum RMS amplitude value for the vowel occurs at time t=T. The range of the signal (RMS or F_1) is 2*D, while the maximum slope is obtained by taking the derivative of the above expression with respect to t, and finding the maximum at t = $\frac{T}{2}$. This maximum slope is D$\frac{\pi}{T}$. Consequently, the slope of the line on a range versus maximum slope plot will be $\frac{2D}{D\frac{\pi}{T}}$, which simplifies to $\frac{2T}{\pi}$. Thus, the time it takes to go from a glide to the vowel is equivalent to the slope times $\frac{\pi}{2}$.

The slope of a given set of data points was obtained by finding the best fit line through the data points, with the constraint that the line goes through the origin. Such a line must pass through the origin because whenever the maximum range is 0, the maximum slope in that region must be zero, and vice versa. Thus, using the familiar mean-square-error criterion on the set of N data points (x_i,y_i) for i=1,2,...,N, the optimal m was sought to minimize $\sum_{i=1}^{N}(y_i - mx_i)$. Taking a derivative with respect to m and setting the result to zero, a closed form solution for m was found. This solution was

$$m = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N}(x_i)^2}. \tag{5.1}$$

Determining Duration for Non-Glides to Vowels

The slopes of the lines fitting the points for the non-glides were also examined. However, the model for these tokens was different from that for a glide. Since non-glides tend to have more abrupt initial transitions, the the region from thirty seconds after the onset of voicing to the vowel was modeled as one fourth of a period of a sinusoidal wave. Thus, having the endpoints being t=0 (thirty seconds from the voicing landmark) and t=T (the location of the vowel), the model for the sinusoid can be written as D $\sin(\frac{\pi t}{2T})$. The range is thus D, and the maximum slope is $\frac{D\pi}{2T}$. Thus, the slope of the line in the range versus maximum slope plot will be $\frac{2T}{\pi}$, as was the result from before.

5.3.2 Differences between VCVs and Sentences

Originally, the VCV database was to be used as the training set, and the sentences database was to be used as the test set. However, it was determined that the rate of speech in isolated utterances (VCV) versus continuous utterances (sentences database) varied enough that the sample means and covariance matrices determined from the VCV database could not be accurately used as standards for hypothesis testing on the sentences database. As a result, the first ten sentences of the sentences database was selected as a means of training and obtaining averages and covariance of the data, while the remaining thirty were used for testing. The following subsections discuss the analysis which led to the realization that the rate of speech in the VCV database was sufficiently different from that in the sentences database for one to be used as a training set and the other to be the test set.

5.3.3 Differences in RMS Amplitude

Figure 5-12 shows a plot of the RMS amplitude range versus the maximum RMS amplitude rate of change for the VCV glides. The following vowel is denoted by the first character (A=/aa/, a=/ah/, E=/eh/, I=/iy/, U=/uw), while the second letter represents the specific speaker is (C=CB, D=DW, K=KS). Contrary to what one might expect, the distribution of data in Figure 5-12 does not conform to the supposition that higher energy vowels should have larger ranges and larger rates of change. A suggested explanation for this, based on listening to the speech represented by the data shown, is that the speakers make an extra tight constriction for the high vowels, since the tongue and vocal tract are already in a narrower constriction for the vowel. Consequently, the range ends up larger, and the rate of change becomes larger to maintain a constant glide to vowel duration.

In the plot, the best fit line is shown, along with an ellipse which represents the equal probability contour at 1.5 standard deviations away from the mean. The outliers, points which fell outside of the ellipse, were looked at individually to see why they deviated from the mean by such a marked amount. It was found that three of the four data points with large rates of change of energy (above .28 dB/msec) were VCVs which had the vowel /iy/. This can be explained by the fact that since the vowel /iy/ already has a somewhat tight constriction in the vocal tract, each speaker must make an even stronger constriction to make the difference from the vowel to the glide and back to the vowel audible. The slope of the best fit line, which represents $\frac{2}{\pi}$ times the time between the glide and the vowel when using a sinusoid as a model, was found to be 42.8 milliseconds. This translates into a glide to vowel time of about **67.2** milliseconds when using the sinusoidal model.

Data for glides in the sentences database were examined next. Figure 5-13 shows a plot of the averages for each of glides in the forty sentences in the sentences database. The slope of the best fit line through the origin for this data was 30.8 milliseconds. Correspondingly, this time showed an average glide to vowel time of **48.4** milliseconds, somewhat smaller than that for the VCVs. The differences in these durations were enough for one to decide that using the VCV database for training would not be adequate for testing on the sentences database. Similar results were obtained for F_1 data. The next section discusses these results.

For each of the non-glide classes, the duration from consonant to vowel was determined. Table 5.3 summarizes the results. As one can see, the durations of non-sonorant consonants are much smaller than those for the glides. Also notice that the unvoiced stops and fricatives are shorter in duration than their voiced counterparts. Such is the case in part because the onset of voicing occurs further in the transition into the vowel for unvoiced segments.

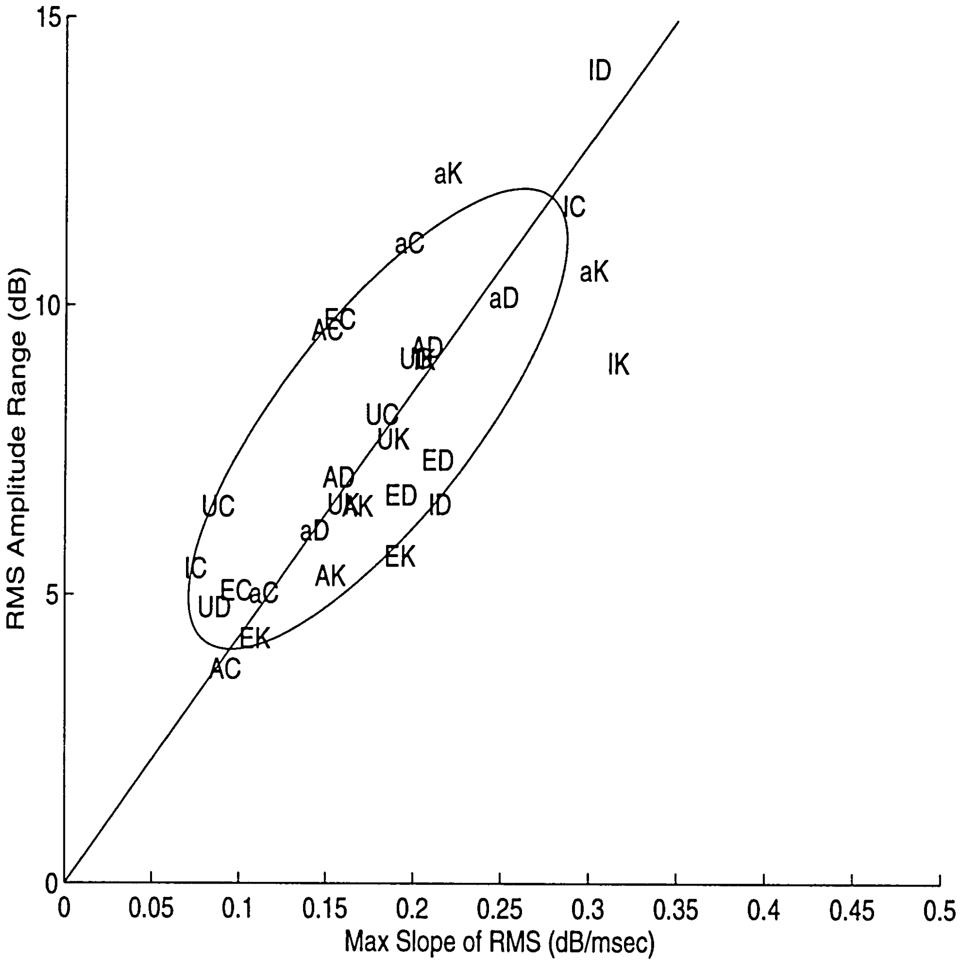

Figure 5-12: Plot of RMS amplitude characteristics of glide tokens from VCV database, along with the 1.5 standard deviation curve and the best fit line through the origin.

Consonant	Duration (msec)
Affricates	26.7
Voiced Stops	20.9
Unvoiced Stops	15.7
Voiced Fricatives	20.4
Unvoiced Fricatives	18.1
Sonorants	
Nasals	62.8
Liquids	68.1
Glides	67.2

Table 5.3: Duration of Different Classes on Consonants as Determined by Fitting the Model for VCV Data (note: the fit on the voiced fricatives included only those values which were obtained from cases where the landmark was correctly detected)

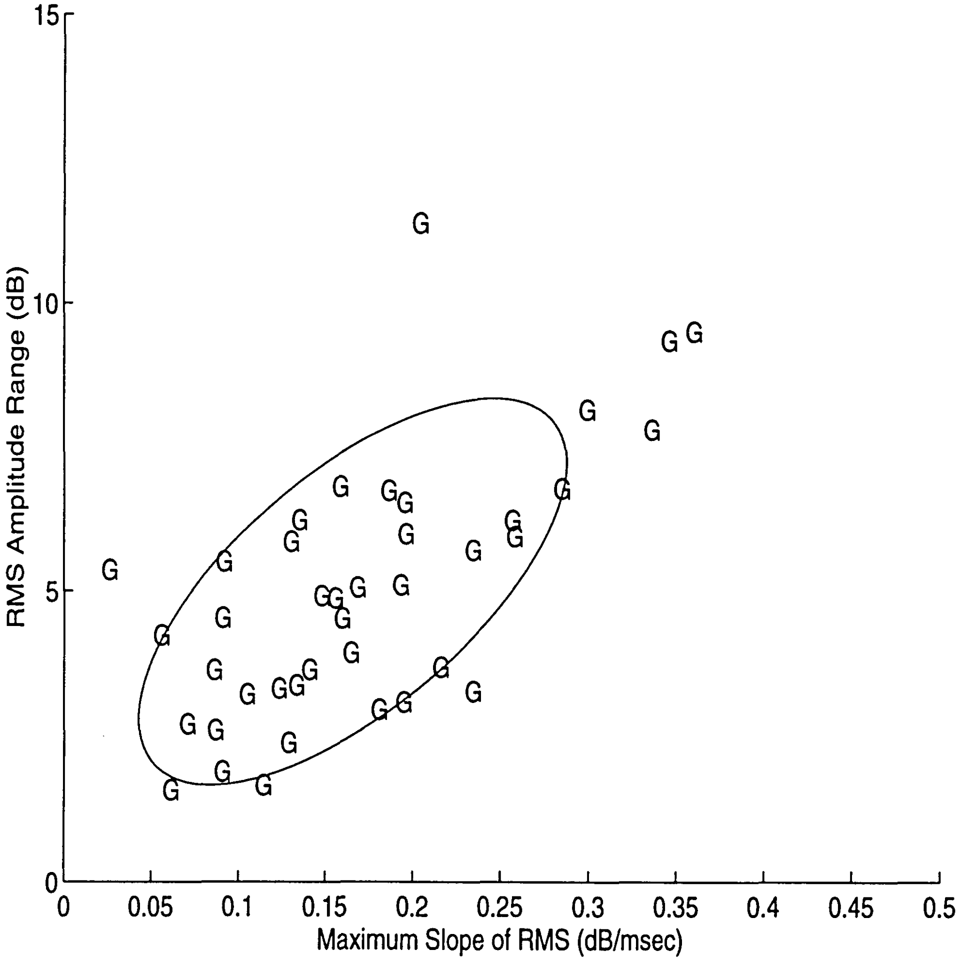

Figure 5-13: Plot of RMS amplitude characteristics of glide from sentences database, along with the 1.5 standard deviation curve. Each data point represents the average of all glides within each of the forty sentences (so, a total of forty data points may be found).

5.3.4 Differences in F_1

As with RMS amplitude, plots of F_1 rates of change versus F_1 range were analyzed. Figure 5-14 shows the distribution of VCV glides on an F_1 range versus maximum F_1 rate of change plot. This plot is noisy because a few major outliers resulted from errors made in the formant tracking algorithm. The slope of the best fit line was found to be 89.0 milliseconds, which translates to a time of **139.7** milliseconds. At first glance, this may seem rather unusual that the time for the F_1 transition is longer than that for the RMS amplitude movement. One may wonder whether the sinusoidal model is still good for F_1. Looking at the plots of transitions in Section 4.2 (Modeling the Transition as a Sinusoid), it can be seen that the transitions for F_1 appear sinusoidal, much like that for the RMS amplitude movements. Thus, this result indicates that it takes longer for the transition of F_1 to settle when going into a vowel, as compared with RMS amplitude. One can explain this in terms of the production model. At the point of maximum constriction in a glide, the vocal tract is so constricted that vibrations in the glottis are affected. Once the constriction becomes less strict, the glottis begins to vibrate normally and at full energy. At this point, the RMS amplitude will have already come close to its final level. However, at this point, the tongue is still in the process of moving to its final vowel position, so the F_1 transition takes more time to complete.

Likewise, the data for the sentences database was computed. The slope of the best fit line was found to be 32.1 milliseconds, which translates to a glide to vowel duration of **50.4** milliseconds. Figure 5-15 shows a plot of the glide distribution, with one point for each of the forty sentences. An average for each of the sentences across all five speakers was made for plotting purposes because plots having all 640 tokens were too congested for reasonable observations to be made. The numerical results detailed later are averages of the total set of tokens, and not an average of these data points in the plots. As in the case of RMS amplitude, the average glide to vowel duration for continuous speech was shorter than that of isolated speech. Interestingly, though, the glide to vowel duration for sentences was more similar between F_1 and

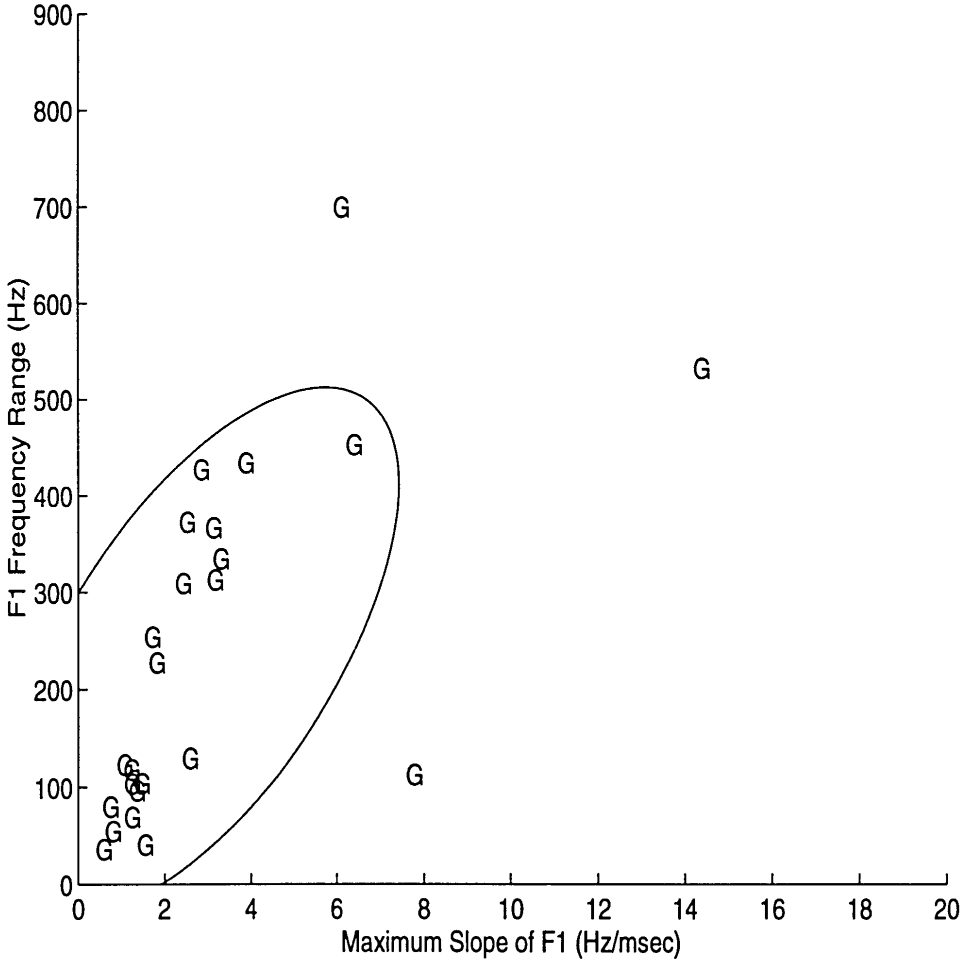

Figure 5-14: Plot of F_1 characteristics of glide tokens from VCV database, along with the 1.5 standard deviation curve.

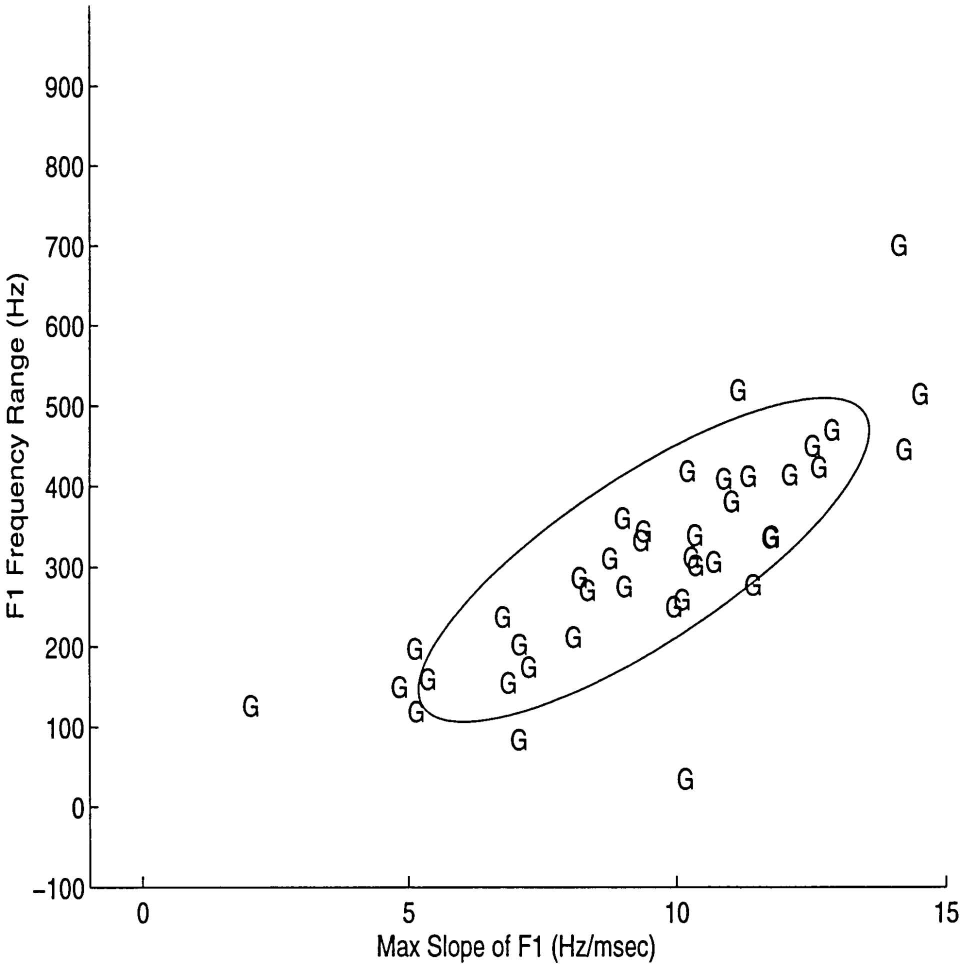

Figure 5-15: Plot of F_1 characteristics of glide from sentences database, along with the 1.5 standard deviation curve. Each point on the graph represents the average of all glide tokens within each of the forty sentences (so, a total of forty points are on the plot).

RMS amplitude than for the same two measurements for the VCV database. A possible explanation is that in continuous speech, the transitions are so rapid that the movements of the vocal tract are less pronounced, thus allowing it to reach the vocal tract's final position about the same time the glottis attains its full vibrating position, instead of being much later in time.

5.4 Summary of Recognition Process

The development of the recognition algorithm was the bulk of the work performed for this thesis. Three important ideas were drawn from this work. First, the fact that a best fit line can yield a reasonably good model of how the data lie allows one to conclude that the duration from the glide to the vowel is more or less constant. As a result, whenever a large range of RMS amplitude or F_1 energy is traversed, the rate of change merely becomes larger.

Second, differences can be seen between the RMS amplitude and F_1 frequency characteristics of glides in isolated versus continuous speech. This fact resulted in the VCV database not being used as a training set. However, the information obtained from the VCV database was useful in learning the relative characteristics of different consonants and following vowels.

Finally, modeling the other consonants in the VCV database allowed one to see how similar the sonorant consonants were to each other, and how different they were as a whole to the other classes of consonants. Voiced fricatives became somewhat of an issue mainly because of errors in the landmark detection and formant tracking. Otherwise, the fricatives, stops, and affricates were clearly separated from the glides.

Chapter 6

Final Algorithm and Results

This chapter summarizes the glide detection results obtained in the VCV database analysis, along with those used for the training and testing performed on the sentences database.

6.1 VCV Results

Hypothesis testing (discussed in Section 2.4) was performed using the measurements of RMS amplitude range, maximum RMS rate of change, F_1 frequency range, and maximum F_1 rate of change. It should be noted that in this case, the testing and training were both the same VCV database, so results will be a little better than in the sentences database case, where the testing and training sets are disjoint. Few errors were made in classification in this database after the landmark errors and F_1 tracking errors were corrected by hand. The results, shown in Table 6.1, account for corrections in F_1 but not the landmark errors. This was done because a better formant tracker may be implemented in the near future, while improvements in the landmark detector may be more difficult because the methods used in this thesis (RMS amplitude, F_1 = low frequency energy) are similar to those used by Liu in her landmark detector. The means and covariance matrices computed were based on the corrected data, since using the errors would not contribute to a better estimate of the actual glide characteristics. Furthermore, during times when the formant tracking

Consonant	Detected as Glide	Detected as Non-Glide
Glides	93.3%	6.7%
Non-Glides	6.6%	93.4%
Divided into classes		
For non-glides		
Liquids	34.6%	65.4%
Nasals	31.6%	68.4%
Affricates	0.0%	100.0%
Voiced Fricatives	23.3%	76.7%
Unvoiced Fricatives	0.0%	100.0%
Voiced Stops	0.0%	100.0 %
Unvoiced Stops	0.0%	100.0 %

Table 6.1: Performance of Thresholds on VCV data; it performs perfectly well on stops, unvoiced fricatives, and affricates

Measurement	Glide	Non-Glide
RMS Max Slope (dB/msec)	.195	.122
RMS Amplitude Range (dB)	8.05	3.29
F_1 Max Slope (Hz/msec)	3.10	4.07
F_1 Frequency Range (Hz)	278.3	163.2

Table 6.2: Average values of the different measurements made on the VCV database.

was difficult to correct, the values found were discarded. The issue of F_1 errors will be discussed in more detail in the future work section. Table 6.1 summarizes the results found for the VCVs. Table 6.2 summarizes the average values found for the different measurements.

6.1.1 Preliminary Sentences Database Issues

As mentioned in the previous chapter, preliminary tests on the sentences database showed that the means of the measurements in the sentences database differed from those of the VCV database. Consequently, the first ten sentences were used as a training set, while the remaining thirty sentences were used as a test set.

A couple of issues arose before any further data could be collected. In the sentences database, finding where to start the analysis was an issue. The basis for the approach taken was twofold. First, as mentioned earlier, it is known that all glides are followed by a vowel. Second, vowels are easier to locate than glides using a formant tracker since the energy of the signal within a vowel is much stronger. Thus, the final problem statement was developed as follows: given the location of vowels and abrupt consonantal releases (as found by Liu's [1995] program), can one determine if a glide exists in the region preceding a vowel (call this the right landmark), and ahead of the previous (left) landmark? Analysis of the waveform data began thirty milliseconds after the landmarks.

The left landmark can be one of two different possibilities. This landmark most frequently is the indicator for the start of voicing. However, when such a landmark is not present, as in the case when a vowel at the end of a syllable is followed immediately by a vowel in the beginning of the next syllable, voicing may never cease. This occurs also when the landmark detector fails to find a landmark between a VCV utterance. When such a situation arises, the most recent landmark before the right landmark will be the previous vowel.

It is known that the transition out of a glide and into the vowel is a smooth one for RMS amplitude. Figure 5-1 shows sample RMS amplitude transitions from a consonant release to a vowel (like from the /s/ to /I/ in *sim*), and from a consonant release through a glide and to a vowel (like from the /s/ to /I/ in *swim*). This model is one which the theory of glides dictates (the slow rise in amplitude, and the small amplitude range).

However, the strong amplitude increase at the point of release became a problem. If analysis began at the release, then sometimes, even if a glide occurs between the consonant and the following vowel, the maximum rate of change would be measured in this initial region, causing to errors in identification. Starting the analysis 30 milliseconds from the landmark solved this problem. Such an adjustment proved to be a major factor in making the analysis of the sentences viable.

6.2 Sentences Database Results

First, averages and the covariance matrix were found for the data collected from the first ten sentences, the training set. Table 6.3 summarizes the averages determined

Consonant	Measurement	Average Value
Glide	RMS Max Slope	.171
Non-Glide	RMS Max Slope	.148
Glide	RMS Amplitude Range	4.95
Non-Glide	RMS Amplitude Range	2.94
Glide	F_1 Max Slope	9.52
Non-Glide	F_1 Max Slope	7.85
Glide	F_1 Frequency Range	318.8
Non-Glide	F_1 Frequency Range	208.5

Table 6.3: Averages found for the glides and non-glides from the first ten sentences.

Consonant type (Speaker)	Classified as Glides	Classified as Non-Glides
Glides (ARI)	86.7%	13.3%
Non-Glides (ARI)	10.5%	89.5%
Glides (JW)	89.4%	10.6%
Non-Glides (JW)	8.1%	91.9%
Glides (KS)	91.5%	8.5%
Non-Glides (KS)	6.3%	93.7%
Glides (MJ)	85.2%	14.8%
Non-Glides (MJ)	11.2%	88.8 %
Glides (SSH)	87.4%	12.6%
Non-Glides (SSH)	10.7%	89.3%
Glides (Total)	88.0%	12.0%
Non-Glides (Total)	9.4%	90.6%

Table 6.4: Overall results found from the sentences database.

in the sentences, while Table B.1 in the appendix lists the covariances of the glides and non-glides. The covariance matrices are left in the Appendix because not much can be gained intuitively from a quick glance at the matrix itself. Its usefulness is primarily in the computation.

Analysis was made on all of the speakers in the thirty remaining sentences. In total, the forty sentences spoken by five speakers totaled 640 glides and 1805 vowels. Of the 1805 vowels, 1522 were full, which 283 were reduced.

Test Results

Table 6.4 tabulates the total results for the different speakers, along with the overall results. From these results, it can be seen that locating non-glides works better than finding glides. However, the percentages for both are still reasonable. Of the non-

Consonant	Percentage of 9.4% Detected as Glide (%)	Overall percentage (%)	Number of Occurrences
/m/	17.7	1.7	11
/n/	16.1	1.6	10
/ng/	8.1	0.8	5
/l/	29.0	2.8	18
/r/	24.2	2.3	15
Voiced Fricatives	3.9	0.5	3
All Others	2.6	0.3	2
Total	100	9.4	62 (out of 640 tokens)

Table 6.5: Breakdown of errors of non-glides being detected as glides. The *overall percentage* indicates the percentage of times that a non-glide, which was classified as a glide, happened to be the particular consonant listed in the leftmost column.

glides which were within the threshold, most of them were sonorants. The next section breaks down these results, and provides a discussion regarding how these errors might be corrected.

Nasals and Liquids being Classified as Glides

As discussed in the previous chapter, nasals, liquids, and glides are all sonorants. Consequently, they share many features. In the analysis, much overlap exists between these three consonant types. Furthermore, the consonant /h/, which can be classified as another glide, fits into this group of consonants as well. From Table 6.4, it can be seen that 9.4% of the time, non-glides were classified as glides. Table 6.5 shows the breakdown of how this 9.4% is distributed. From this table, it can be seen that the problem lies mostly in these other sonorants and a little bit with the voiced fricatives.

The errors attributed to something other than a sonorant or a voiced fricative can be assumed to be an esoteric error, possibly caused by either errors in labeling, or an unusual variability in the speech. For the others, however, the results are reasonable. To remove the errors due to the nasals, Liu's program was run again to find the abrupt landmarks in nasals. The landmark detector written by Liu locates abrupt consonants. The algorithm designed by her also has the capabilities of finding the abruptness of nasals most of the time when abruptness in the mid-frequency range is measured.

Sonorant Landmark Detector Errors

Liu's landmark detector finds three types of landmarks, **g**, **s**, and **b**, where g=glottis, s=sonorant, and b=burst. The glottis landmark determines the time in which the amplitude of glottal vibration changes abruptly. This is the landmark used to find the release point of abrupt consonants. The sonorant landmark supposedly finds closures and releases of nasals and the /l/. To find these landmarks, the routine searches for high frequency abruptness in energy bands from 1.3 kHz to 8 kHz. Closures created in the nasals or the /l/ cause quick changes in higher frequency energy [Liu, 1995].

This algorithm was tested on some of the sentences in the sentences database. Unfortunately, it was found that the accuracy of this measure was not very good. In fact, for a set of ten randomly selected sentences in the database of the five speakers, it was found that 50% (35 out of 70) of the actual glides that were found by Liu's landmark detector were erroneously labeled as abrupt sonorants (nasals or the /l/). Meanwhile, 43% (26 out of 61) occurrences of abrupt sonorants were not identified as sonorants landmarks. With accuracy around 50%, using this information would likely not help in increasing recognition accuracy. In essence, little or no improvement can be had from using this landmark detector to filter out nasals and liquids from the glides. Consequently, the issue of finding a good detector for nasals is a possible topic for future research.

Effect of Context on Glides

Preliminary work by the author [Sun, 1995a] showed that the context of a glide within a sentence had an effect on its characteristics. Consequently, while analyzing the data from the sentences database, the effect of stress was considered. In many situations, normal, fluid speech introduces contractions or simplications of spoken utterances. For example, sentence six is *Blow them, and the leaves will fly away.* For all five speakers, the /w/ in will was deleted. Consequently, the sentence became *Blow them, and the leaves'll fly away.* Deleted glides were not considered as glides in the analysis.

In the same sentence, the vowel in the word *the* was reduced. Such a reduction

Consonant type	Classified as Glides	Classified as Non-Glides
Reduced Vowels		
Glides	76.7%	23.3%
Non-Glides	13.7%	86.3%
Full Vowels		
Glides	90.1 %	9.9%
Non-Glides	8.7%	91.4 %

Table 6.6: Comparison of Glide Recognition Results for Consonants Followed by Full Vowels and those by Reduced Vowels.

shortened the length of the utterance, and also kept the amplitude of the vowel low compared to other vowels in the same sentence. These characteristics are normal for reduced vowels. The decrease in energy poses more of a problem for the glide detection algorithm because the change in amplitude becomes smaller, thus moving towards the mean of measurements consistent with non-glides (which would lead to a hypothesis test result of a non-glide). Table 6.6 shows the performance of the algorithm on consonants followed by reduced vowels compared with the performance on full vowels.

From this table, it can be observed that for full vowels, the recognition rate (number of times glides are properly classified as glides plus number of times non-glides are properly classified as non-glides divided by the total number of vowels analyzed) is about *90.8%*. Of the incorrectly labeled non-glides, many of the errors come from other sonorants which have characteristics similar to those of glides. Once a better detector for these consonants is found, the overall accuracy rate of this program can be improved.

Common Source of Errors

Mistakes in glide detection seemed to stem from a few isolated sources. As described already, the problem of detecting nasals and liquids as glides was a common problem. In addition, voiceless glides, as in *which*, were difficult to locate. This was the case because, at times, the landmark indicating the onset of voicing was beyond the location of the glide. Consequently, the search for the glide was begun in a region past the actual location of the glide landmark. Errors in precision from the landmark detector and from manually labeling the location of the vowels also probably contributed to the overall errors.

Chapter 7

Conclusions and Future Work

The approach of finding glides before vowels is one which yields good results in continuous speech. The algorithm developed used knowledge about glides and other sounds in the English language, along with results from previous results in knowledge-based speech recognition and lexical access ([Liu, 1995], [Espy-Wilson, 1993], and [Li, 1993]). Previous work detected glides in isolated carrier phrases, and labeled abrupt consonants. This work combined these two ideas, and detected glides in clearly spoken sentences much in the same way that abrupt consonantal landmarks were found.

7.1 Summary

The process of developing the algorithm for detecting glides eventually required a search of literature describing the attributes of all consonants in the English language. At first, it was believed that just understanding the characteristics of glides was all that was necessary to locate them in speech. A literature review of Stevens [forthcoming], Arman and Stevens [1993], and other papers provided enough information for a reader to know that /w/'s have extremely low F_1 and F_2, while /j/'s have extremely low F_1, and extremely high F_2. Furthermore, one would also know that the RMS amplitude in glides is weaker than that in vowels. However, the problem is that other sonorant consonants exhibit similar reductions in RMS amplitude. For example, the liquid /r/ has a low F_2, also. So, an analysis of nasals, liquids, and obstruent consonants was made, and findings were briefly summarized in chapter two.

A new database of sentences was created which was full of glides. Originally, it was proposed that the VCV database be used as the training data, and then the sentences as the test data. However, a thorough analysis of the two databases showed differences between isolated (VCV) and continuous (sentences) speech; so, instead, one-fourth of the sentences database was used as the training set, while the remaining 75% served as the test set. The measurements collected included maximum RMS amplitude change, RMS amplitude range (from glide to vowel), maximum F_1 frequency change, and F_1 frequency range. The training data were used to obtain sample means and covariances for glide and non-glide tokens. Then, a hypothesis test was used for each new set of measurements made on a test token, and a determination was made whether the token was a glide or not. Implicit in this analysis was the fact that the duration from the release to the center of the following vowel had to be at least 30 milliseconds in order for the region to even be tested for the existence of a glide. So, for durations less than 30 milliseconds, these tokens were automatically labeled as non-glides.

It was determined that potential problems could occur with nasals, liquids, and voiced fricatives. Knowledge about nasals and liquids already made it reasonable for one to believe that the similarities of these sonorants would made it difficult to separate glides from the nasals and liquids. This algorithm was performed on the VCV database, with the resulting glide detection rate of 93.3%, and the non-glide detection accuracy of 93.4%. For non-glides which were classified as glides, none of these errors were attributed to stops or unvoiced fricatives, which was reasonable based on the vast differences in the production models of these consonants compared with glides. The problems were more with the nasals and liquids.

When this algorithm was tested on the test section of the sentences database, glides were correctly detected 88.0% of the time, and non-glides were found as such 90.6% of the time. These results were similar to the ones found for the VCVs, but somewhat worse. As expected, 91.9% (57 out of 62) of the non-glides which were found to be glides were actually nasals or liquids. With sentences, it was possible to look at glides and non-glides before reduced vowels. Results were somewhat worse under these conditions, with glide detection rates at 76.7%, and non-glides being found 86.3% of the time, as opposed to 90.1% of the time for glides before full vowels, and 91.4% for non-glides before full vowels.

Many different approaches can be made towards recognizing glides, but as of now, none of them provide an error-free way of identifying all glides in speech, especially in continuous speech where vowels are sometimes reduced, and words occasionally not clearly spoken. Root-Mean-Square amplitude seems to be a useful technique because theory clearly indicates that the tight constriction of glides causes strong decrease in energy. Likewise, though, formant values and transitions, when correctly tracked, provide excellent information in finding glides because of their extreme values.

7.2 Future Work

Much future work is needed to develop landmark detection for the complete lexical access project. The next immediate step is locating the vowels from formant and RMS amplitude information. This is of considerable importance because the work presented here uses the assumption that the vowel landmarks will soon be found. As mentioned in earlier chapters, this assumption is reasonable because formant trackers are generally reliable in regions of high energy, like vowel regions. Analysis of particular energy bands, as previous work has done, may provide additional information uniquely characterizing different phonemes in the English language.

A nasal and liquid detector is also another useful project for future work. Such a detector would have immediate effect on the work in this thesis, as one of the problems of glide detection involves finding and removing nasals and liquids from the potential set of glides detected. An approach at finding the nasals should probably take advantage of the known fact that the spectrum has an extra pole/zero pair, which appears because of the nasal cavity opening. It would appear as if locating nasals would not be so difficult; with the strong dip in amplitude at a low frequency (around 1100 to 2000 Hz) due to the zero from the nasal cavity opening. Liquids, on the other hand, may be more difficult to detect. The problem of finding them is similar to the difficult problem of finding glides without any other landmarks.

A locally adaptive method may be considered to locate glides. This method, which was looked at in Sun [1995b], allows the detection of glides without the need of finding the location of the following vowel. The major drawback of this analysis was the lack of a highly effective formant tracker. In fact, in the analysis made within this thesis, formant tracking errors occurred in the glide region (region where measurements were important for correct recognition) about 6% of the time. Although this number may seem small, such errors can lead to the degradation of detection by that amount, and a reduction of accuracy from, say 94% down to 88% would be a significant decrease in correct recognition. Improvements in formant tracking may make this technique viable in the future.

The idea of curve-fitting glide transitions with sinusoids is also a promising area of research. Some researchers today are using sinusoids to model transitions in synthesis projects. The smooth transition seems to naturally imply such a model, so work in synthesis may bring to light some more possibilities for modeling glides in the future.

7.3 Bibliography

Arman, G. and K.N. Stevens (**1993**). *Acoustic characteristics of glides*, Speech Communication Group, MIT; Unpublished Manuscript.

Bickley, C.A. and K.N. Stevens (**1986**). *Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies*, **J. of Phonetics 14**, pp. 373-382.

Chiba, T. and M. Kajiyama (**1941**). **The Vowel-Its Nature and Structure**, Tokyo, Japan, p. 188.

Duda, R.O. and P.E. Hart (**1973**). **Pattern Classification and Scene Analysis**, John Wiley and Sons, New York.

Entropic Signal Processing System (ESPS) (**1992**). Software speech processing package.

Espy-Wilson, C.Y. (**1987**). *An acoustic-phonetic approach to speech recognition: Application to the semivowels*, **Massachusetts Institute of Technology, PhD Thesis**.

Fant, G. (**1959**). *Acoustic analysis and synthesis of speech with applications to Swedish*, **Ericsson Technics 15**, pp. 3-108. [Data used from this found in Fant [1973] Ch. 3).

Fant, G. (**1973**). **Speech Sounds and Features**, Cambridge, MA, pp. 36-37, 84-93.

Glass, J.R. (**1988**). *Finding acoustic regularities in speech: applications to phonetic recognition*, **Massachusetts Institute of Technology, PhD Thesis**.

Goldstein, U.G. (1980). *An articulatory model for the vocal tracts of growing children*, **Massachusetts Institute of Technology, PhD Thesis.**

Hayes, M.H. (1996). **Statistical Digital Signal Processing and Modeling**, John Wiley and Sons, New York.

Johnson, M.A. (1996). *Formant and burst measurements with quantitative error models for speech sound classification*, **Massachusetts Institute of Technology, PhD Thesis.**

Johnson, R.A. and D.W. Wichern (1982). **Applied Multivariate Statistical Analysis**, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Lehiste, I. and G.E. Peterson (1961). *Transitions, glides, and diphthongs*, **J. Acoust. Soc. Am. 33**, pp. 268-277.

Li, P. (1993). *Feature modifications and lexical access*, **Massachusetts Institute of Technology, Bachelor's Thesis.**

Liu, S.A. (1995). *Landmark detection for distinctive feature-based speech recognition*, **Massachusetts Institute of Technology, PhD Thesis.**

Loftus, G.R. and E.F. Loftus (1988). **The Essence of Statistics**, McGraw-Hill, Inc., New York, 2nd edition.

Mack, M. and S.E. Blumstein (1983). *Further evidence of acoustic invariance in speech production: The stop-glide contrast*, **J. Acoust. Soc. Am 73**, pp. 1739-1750.

Manly, B.F.J. (1986). **Multivariate Statistical Methods: A Primer**, Chapman and Hall, New York, "Ch. 7 Discriminant Function Analysis".

Miller, J.L., and T. Baer (**1983**). *Some effects of speaking rate on the production of /b/ and /w/*, **J. Acoust. Soc. Am. 73**, pp. 1751-1755.

Moser, H., J.J. Dreher, and H.J. Oyer (**1957**). **One-Syllable Words.**, Technical Report 41, The Ohio State University Research Foundation.

Niblack, W. (**1986**). **An Introduction to Digital Image Processing**, Prentice Hall, pp. 115-116.

Oppenheim, A.V. and R.W. Schafer (**1989**). **Discrete-Time Signal Processing**, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Peterson, G.E. and H.L. Barney (**1952**). *Control methods used in a study of the vowels*, **J. Acoust. Soc. Am. 24**, pp. 175-182.

Rabiner, L.R. and R.W. Schafer (**1978**). **Digital Processing of Speech Signals**, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Stevens, K.N. (**Forthcoming**). **Acoustic Phonetics.**

Sun, W. (**1995a**). *Effect of context on formant values of glides*, **6.542J Laboratory on the Physiology, Acoustics, and Perception of Speech**, MIT, Fall 1995, Term Project, Unpublished Manuscript.

Sun, W. (**1995b**). *Study of glide features that assist in the automatic speech recognition process.*, **Research Summary**, MIT, Fall 1995, Unpublished Manuscript.

Talkin, D. (**1987**). *Speech formant trajectory estimation using dynamic programming with modulated transition costs*, **J. Acoust. Soc. Am. Suppl. 1, 82**, p. S55.

Willsky, A.S., G.W. Wornell, and J.H. Shapiro (**forthcoming**). *Hypothesis Testing*, Supplementary Notes for MIT Course 6.432, Stochastic Processes, Detection, and Estimation (Chapter 2).

Appendix A

MATLAB M-file Code

```
function out = computerms(in,samp,windlen)
% COMPUTERMS    Computes the rms of a speech signal
% Usage:        OUT = COMPUTERMS(IN,SAMP,WINDLEN)
%               IN is the input waveform, SAMP is the
%               sampling rate of the waveform, and
%               WINDLEN is the length of the window

% WSun 4/15/96

in = in(:);
wind = ceil(windlen*samp); % Create window length
% Zero pad at both ends
in = [zeros(ceil(windlen*samp/2),1);in;zeros(ceil(windlen*samp/2),1)];
leng = length(in);
out = zeros(leng-wind+1,1);
% Compute RMS at each point
for ctr = 1:(leng-wind+1),
  out(ctr) = sqrt(sum(in(ctr:ctr+wind-1).^2)/wind);
end
out = 20*log10(out);
```

Table A.1: MATLAB M-file for ComputeRMS.m; program to compute RMS Amplitude.

```
function [A,B,omega,MSE,xhat,x] = bestsine(x,vowel);
% BESTSINE  [A,B,omega,MSE,xhat] = bestsine(x,vowel)
% where the best sine fit is:  xhat = A - B cos(omega t)
% x is the data input and vowel is the location of the vowel (samplewise)
% This program assumes a data sampling rate of 500 Hz.

% WSun 5/6/96

% Smooth data
x = x(:);
x = conv(x,ones(5,1)/5); % Smooth via 10 msec window
Len = length(x);
x = x(5:Len-4); Len = Len - 8;
[tmp,a] = min(x);
data = x(a:vowel);
newlen = length(data);
val = [0.002*(1-newlen):0.002:0];
freq = 2*pi*[(1/.35):.01:(1/.15)]; % Frequency Range
coswt = cos((val')*freq); % Matrix of values for cos(wt) evaluation
xcos = (data')*coswt;
pcos = sum(coswt);
sqcos= sum(coswt.^2);
datasum = sum(data);

% Solve for parameters as a function of frequency
Bx = ( (newlen*xcos) - datasum*pcos )./( pcos.^2 - newlen*sqcos );
Ax = (Bx.*pcos + datasum*ones(1,length(freq)))/newlen;

error = ones(newlen,1)*Ax - (coswt).*(ones(newlen,1)*Bx) - ...
  data*ones(1,length(freq));
MSEE = sum(error.^2);
[value,loc] = min(MSEE); % Find frequency that minimizes MSE
omega = freq(loc);
MSE = MSEE(loc);
B = Bx(loc);
A = Ax(loc);
xhat = A - B*cos([-.125:.001:0]*omega);
```

Table A.2: MATLAB M-file for BESTSINE.M; program to find best sine fit.

Appendix B

Covariance Matrices

$$C_{glides} = \begin{pmatrix} .005 & .149 & .015 & -1.531 \\ .149 & 7.095 & 1.755 & 22.914 \\ .015 & 1.755 & 8.490 & 421.8 \\ -1.531 & 22.914 & 421.8 & 30750 \end{pmatrix}$$

Table B.1: Covariance matrix for glides (with variables in the following order: RMS Max Slope, RMS Amplitude Range, F_1 Max Slope, F_1 Frequency Range).

$$C_{non-glides} = \begin{pmatrix} .005 & .1277 & .006 & -1.815 \\ .1277 & 8.25 & 3.05 & 92.07 \\ .006 & 3.05 & 9.08 & 450.6 \\ -1.815 & 92.07 & 450.6 & 31355 \end{pmatrix}$$

Table B.2: Covariance matrix for non-glides (with variables in the following order: RMS Max Slope, RMS Amplitude Range, F_1 Max Slope, F_1 Frequency Range).