ER ENGINEERING LIBR

TK7855
.M41
.R43
no.531

MASS. INST. TECH.
MAR 16 1988
LIBRARIES

An Acoustic-Phonetic Approach to Speech Recognition: Application to the Semivowels

RLE Technical Report No. 531

June 1987

Carol Yvonne Espy-Wilson

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

AN ACOUSTIC-PHONETIC APPROACH TO SPEECH RECOGNITION: APPLICATION TO THE SEMIVOWELS

by

Carol Yvonne Espy-Wilson

B.S., Stanford University
(1979)

S.M., Massachusetts Institute of Technology
(1981)

E.E., Massachusetts Institute of Technology
(1983)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 22, 1987

©Carol Yvonne Espy-Wilson

The author hereby grants to MIT permission to reproduce and to distribute copies of this thesis document in whole or in part.

Signature of Author Carol Y. Espy-Wilson
Department of Electrical Engineering and Computer Science

Certified by Kenneth N. Stevens
Kenneth N. Stevens
Thesis Supervisor

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

AN ACOUSTIC-PHONETIC APPROACH TO SPEECH RECOGNITION: APPLICATION TO THE SEMIVOWELS

by

Carol Yvonne Espy-Wilson

Submitted to the Department of Electrical Engineering and Computer Science on May 22, 1987 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

A framework for an acoustic-phonetic approach to speech recognition was developed. The framework consists of: 1) specifying the features needed to recognize the sounds or class of sounds of interests; 2) mapping the features into acoustic properties based on relative measures so that they are relatively insensitive to interspeaker and intraspeaker differences; 3) developing algorithms to extract automatically and reliably the acoustic properties; and 4) combining the acoustic properties for recognition.

The framework was used in the development of a recognition system for the class of English sounds known as the semivowels /w,y,r,l/. Fairly consistent recognition results were obtained across the corpora used to develop and evaluate the semivowel recognition system. The corpora contain semivowels which occur within a variety of phonetic environments in polysyllabic words and sentences. The utterances were spoken by males and females who covered eight dialects. Based on overall recognition rates, the system is able to distinguish between the acoustically similar semivowels /w/ and /l/ at a rate better than chance. Recognition rates for /w/ range from 21% (intervocalic context) to 80% (word-initial context). For /l/, recognition rates range from 25% (prevocalic context following an unvoiced consonant) to 97% (sonorant-final context). However, if lumped into one category, overall recognition rates for these semivowels range from 87% to 95%. Consistent overall recognition rates around 90% were obtained for /r/ and overall recognition rates in the range 78.5% to 93.7% were obtained for /y/.

Several issues were brought forth by this research. First, an acoustic study revealed several instances of feature assimilation and it was determined that some of the domains over which feature spreading occurred support the theory of syllable structure. Second, an analysis of the sounds misclassified as semivowels showed that, due to contextual influences, the misclassified vowels and and consonants had patterns of features similar to those of the assigned semivowels. This result suggests that the proper representation of lexical items may be in terms of matrices of binary features as opposed to, or in addition to, phonetic labels. Finally, the system's recognition of semivowels which are in the underlying transcription of the utterances, but were not included in the hand transcription, raises the issue of whether hand-transcribed data should be used to evaluate recognition systems. In fact, it appears as if insights into how speech is produced can also be learned from such "errors."

Thesis Supervisor: Kenneth N. Stevens
Title: Clarence J. LeBel Professor of Electrical Engineering

DEDICATION

To mom and John

Acknowledgements

Attempting to thank all of those who, second to God, helped me along the way is impossible, for the list is far too long. For those who are left out, I thank you first for all the assistance you gave.

I extend my deepest gratitude to Ken Stevens, my thesis advisor, who has played an enormously helpful role in this endeavor. His guidance and ability to get at the basic and fundamental issues, tempered with his compassion, has made this research uniquely rewarding and pleasurable. I will always be grateful to Ken for his academic and personal concern, especially during the early years of my doctoral studies.

I also thank my readers John Makhoul, Joe Perkell, Stephanie Seneff and Victor Zue. Their enthusiasm towards this research was very encouraging, and their insights, guidance and critiques of versions of this thesis are most appreciated.

Many thanks to the Xerox Corporation, Dr. Herb James in particular, for financial support; and to DARPA for use of the facilities that they provided.

Thanks to NL, SM, MR and SS for serving as subjects.

To past and present members of the Speech Communication Group who provided an intellectually stimulating and pleasant environment, I say thank you. Among this group are several persons whom I specially thank for the many discussions in which we sought to understand some difficult issues. They include Corine Bickley and Stephanie Shattuck-Huffnagel. Along this line, there are others who carefully read parts of this work and offered several helpful suggestions. They include Abeer Alwan, Caroline Huang, Hong Leung, Mark Randolph and John Pitrelli. To them, I extend my gratitude.

In addition, I thank Victor Zue for providing the LISP machines and I thank all of those who have developed the software tools which make this facility an excellent research environment. Finally, for their assistance with these and other tools, I thank Scott Cyphers, Rob Kassel, Niels Lauritzen, Keith North and Mark Randolph.

I am profoundly grateful to several special friends for their support. I will not list them all. However, I must acknowledge Oliver Ibe, John Turner (Dean of the Graduate School at MIT) and Karl Wyatt who were of tremendous help during some critical times; and Austin and Michelle Harton who always had time to listen and care.

My mother and three brothers have always been a source of strength. I cannot find words which adequately express my gratitude to them for their belief in me and their unyielding support.

Finally, I thank my husband, John. His patience, encouragement and counsel were boundless. To him, I am deeply indebted. As John once said of me, he is, very simply, love.

Biographical Note

Carol Yvonne Espy-Wilson was born in Atlanta, Georgia, on April 23, 1957. She received her B.S. in Electrical Engineering in 1979 from Stanford University. Upon leaving Stanford, she came to MIT where she received her M.S. in Electrical Engineering and Computer Science in 1981. The Master's thesis is titled "The Effects of Additive Noise in Signal Reconstruction from Fourier Transform Phase." Realizing she had long since fulfilled the requirements for an E.E. degree, she applied and received it in 1984. In the fall of 1981, she became a member of the Speech Communication Group under the guidance of Dr. Kenneth Stevens. Carol has had several opportunities to present her work at conferences and she is a member of Sigma Xi and other professional societies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Introduction

The ultimate goal of most speech recognition research is the development of a system which allows the natural communication by speech from people to machines. That is, we want recognition systems to be capable of understanding fluent conversational speech from any random speaker. Such systems are desirable since speech is our most natural mode of communication. Thus, unlike today when people must have special skills such as typing to communicate with a computer, the use of such recognition systems requires no training. Furthermore, since we speak much faster than we write and type, speech provides the highest potential capacity in human-to-machine communication. Finally, computers which understand speech free the eyes and hands of the operator to perform other tasks simultaneously.

Although research in speech recognition and other related areas has been going on for several decades, recognition systems have yet to come close to realizing their full potential. With current systems, reasonable recognition performance is possible only if the task is greatly simplified. Present state-of-the-art systems, with few exceptions, can only recognize a small vocabulary of acoustically distinct words which must be said in isolation by a particular speaker. Systems capable of understanding continuous speech also reduce the recognition task by limiting the user to a particular speaker and by constraining the way in which sentences can be formed.

One major reason for these necessary limitations is our present inability to deal with the considerable variability in the speech signal. In addition to linguistic information, the speech signal contains extralinguistic information regarding the talker's personal characteristics, his or her psychological and physiological state, and the recording environment. Thus, to achieve the goal of speaker-independence and continuous speech input, recognition systems must be able to separate out and decode the message-bearing components of the spoken utterance.

What are these message bearing components? We believe that the answer to this question is based on two widely accepted premises. First, the speech signal is composed of a limited set of basic sound units known as phonemes. In English, the inventory of phonemes includes about 16 vowels and 24 consonants. Second, the canonic representation of each phoneme is characterized by a small set of distinctive features, where a distinctive feature is a minimal unit which distinguishes between two maximally close but linguistically distinct speech sounds. For example, the single feature *voice* separates the phonemes /b/ and /p/. The distinctive features also organize the speech sounds into natural classes on the basis of common characteristics. For example, the feature *nasal* lumps the phonemes /m/, /n/ and /ŋ/ into one such class. In languages in general, there are about 20 distinctive features. However, any one language only uses a subset of 10 to 15 for signaling phonetic contrasts.

Although the associations are not well understood in every case, it is hypothesized that all the distinctive features have acoustic correlates. While the distinctive features are binary in nature, the corresponding acoustic properties can have varying degrees of strength due to the wide variability in the acoustic realization of the phonemes. This variability is principally of two types. As we stated earlier, one kind of variability is due to the different vocal tract sizes and shapes of different talkers and the changes in voice quality within the same speaker and across speakers. While there are definite acoustic changes due to these sources, the feature specification of the phonetic segments is usually unchanged. Thus, if properly defined, acoustic properties for features should not be affected by such variability.

On the other hand, another kind of variability known as feature assimilation can modify considerably the feature make-up of the underlying phonemes and the strength of their corresponding acoustic properties. These changes, which occur when phonemes are concatenated to form larger units such as syllables, words and sentences, are due in part to the varying degrees of sluggishness in the articulators when moving from one target configuration to the next. That is, the adjustment of the articulators to implement one set of features may be influenced by the adjustment needed to produce an adjacent set. As a consequence, one or more features of a phonetic segment may spread to a nearby sound, resulting in several types of modifications.

First, some of the features of a segment may change. For example, this phenomenon will sometimes occur when a weak voiced fricative (/v/ and /ð/) is in an intervocalic position. Whereas fricatives are characteristically ***nonsonorant*** with some high frequency noise, in this context they can be ***sonorant*** with no noise. However, features other than the ***sonorant*** feature remain unchanged. Such variants from the canonical representation of a particular phoneme are referred to as allophones. Thus, a /v/ which occurs between two vowels is usually a different allophone from the one which occurs in other contexts. Second, a feature which is normally unspecified for a segment may become specified. An example of this phenomenon is the nasalization of vowels when they are adjacent to a nasal consonant. Finally, a result of this feature spreading may be the merging of two segments into one segment which has a number of features common to both of the underlying sounds. This phenomenon is often seen at word boundaries in continuous speech. For example, the word pair "did you" is often pronounced in fluent speech as "dija." That is, the word-final /d/ and the word-initial /y/ can be coarticulated such that the resulting sound is a /ǰ/. The degree to which sounds undergo such feature assimilation is determined by several factors such as speaking style, speaking rate and language specific rules.

Thus, the use of phonetic features as basic units upon which larger units such as phonetic segments, syllables, words, sentences, etc. are recognized is appealing since, if properly defined and extracted, they should not be affected by much of the within-speaker and across-speaker variability seen in the speech signal. However, it appears that some mechanism is needed to account for feature assimilation effects.

Before outlining and discussing these issues within the context of the class of sounds focused upon in this thesis, we will first consider previous work in speech recognition. A brief review of some of the findings of previous acoustic and perceptual studies of the semivowels, along with the results of an acoustic study conducted in this thesis, are given in Chapter 3.

1.2 Literature Review

Considerable effort has been expended in the development of isolated word and continuous speech recognition systems. Basically, there have been two standard approaches: phonetically-based methods and mathematically-based models.

The phonetically-based approach to speech recognition has mainly been pursued in academia because of its long term investment. This method draws on the distinctive feature theory first proposed by Jakobson, Fant and Halle (1952) and later expanded by Chomsky and Halle (1968). Such an approach attempts to extract the message-bearing components of the utterance explicitly by extracting relevant acoustic properties. While this approach has a strong theoretical base, limited success has been obtained because of the lack of a good knowledge of acoustic phonetics and other related areas. That is, researchers have yet to uncover the proper acoustic properties for features and, therefore, they have not been able to reliably extract this information for phonetic recognition. In addition, all aspects of feature assimilation are not understood.

Researchers of the mathematically-based methods find the well-defined algorithms which can be used within this framework attractive, and many consider the heuristics used in the extraction of explicit speech knowledge ad hoc. This type of an approach to speech recognition has mainly been pursued in industry because of its near term success for constrained recognition problems. Such an approach attempts to extract the message-bearing components of the utterance implicitly. That is, equipped with large amounts of training data and sophisticated engineering techniques, recognition systems are expected to either discover all of the regularities in the speech signal and "average out" all of the variability, or effectively model all of the variability. Presently, none have been able to adequately cope with all types of variability.

Because of the shortcomings of the mathematically-based approaches and yet their ability to model some speech variability that we presently do not understand, there have been recent efforts to develop ways of incorporating our increasing acoustic phonetic knowledge within the statistical frameworks. It is hoped that such an integration of approaches will eventually lead to speaker-independent continuous speech recognition.

In this section, we give a brief review of these methods. For a more extensive coverage of speech recognition research, we recommend reviews given by Lindgren (1965) and Lea (1980).

1.2.1 Phonetically-Based Approach

Recognition systems which attempt to extract acoustic cues from which phonemes or phones are recognized date as far back as 1956 when Wiren and Stubbs developed a binary phoneme classification system. In this system, acoustic properties were used to classify sounds as *voiced-unvoiced*, *turbulent-nonturbulent*, *acute-grave*, and *compact-diffuse*. Although no overall recognition score is given, the performance of this system is encouraging in light of how little was known in the area of acoustic phonetics at the time of its development. For example, vowels in monosyllabic words spoken three times each by 21 talkers were correctly classified as *acute* or *grave* 98% of the time.

Since that time, several recognizers based on this approach have been developed. While most of these systems have obtained only moderate recognition rates for a particular class of phonemes occurring in specific contexts, important concepts have been introduced. For example, Martin, Nelson and Zadell (1964) used detectors which not only indicated when a feature was present or absent, but also indicated the strength of its acoustic correlate. As another example, Medress (1965), as far as we know, was the first to take advantage of phonotactic constraints which restrict allowable phoneme sequences. This information was used to help identify word-initial and word-final consonant clusters in an isolated word recognition system.

More recently, this approach has been applied to the recognition of continuous speech. Between 1971 and 1976, the Advanced Research Projects Agency (ARPA) funded the largest effort yet to develop continuous speech recognition systems. (See Klatt (1977) for a review.) While these systems used some knowledge of acoustic phonetics, most of them relied extensively upon high level knowledge of syntax and semantics for sentence decoding. For example, Harpy, the most successful system in terms of word and sentence accuracy, correctly recognized 97% of the words in the utterances even though it correctly recognized only 42% of the phonetic segments. This poor phonetic recognition was due to a primitive front end which segmented and labelled the speech signal. Whereas the segmenter used acoustic cues extracted from parameters such as zero crossing rates and smoothed and differenced waveforms, the labeller used phone templates consisting of linear-prediction spectra. To deal with variability due to feature assimilation, 98 templates were used to represent all possible allophones, and juncture rules accounted for some effects between phone sequences. In addition, to deal with within-speaker variability, each template was computed by averaging all occurrences of the particular allophone in a set of training sentences.

An exception to this heavy reliance on high level knowledge for continuous speech recognition was the HWIM system developed at BBN which used considerably more acoustic phonetic knowledge. To provide a phonetic transcription of an utterance, a parametric representation and a set of 35 ordered acoustic-phonetic rules was used. This processing resulted in a segment lattice which provided multiple segmentation paths for portions of an utterance. With a dictionary of 71 allophones, 69% of the correct phonetic segments were in the top two choices produced by the front end.

1.2.2 Mathematically-Based Methods

Most commercially available speech recognition systems are based on general pattern-matching techniques which use little speech-specific knowledge. They are speaker dependent and recognize a limited vocabulary of words which must be said in isolation. These systems are trained by having the talker to be recognized generate a set of reference patterns or templates which are digitized and stored. The templates usually consist of a series of spectral sequences computed every 10 to 20 msec. For recognition, these systems use a distance metric to select from a set of stored templates the closest match to the pattern computed from the incoming word. The first complete recognizer of this sort was developed in 1952 by Davis, Biddulph and Balashek. This speaker-dependent system had a recognition rate of 97% for the digits zero(oh) to nine.

Since that time, several engineering techniques have been introduced to deal with some of the variability in the speech signal. For example, to deal with varying speaking rates which result in duration differences between stored and input templates, several time-alignment procedures have been developed. Presently, the most effective and widely used technique is dynamic time warping (DTW), introduced by Sakoe and Chiba (1971). This algorithm, when comparing two templates, uses a distance metric to nonlinearly warp the time axis of one so that the templates are maximally similar. A computationally efficient distance metric developed for use with DTW was developed by Itakura in 1975.

In addition, since spectral templates are inherently speaker dependent, techniques have been developed so that systems could accommodate multiple speakers. One such system, developed by Rabiner et al. (1979), uses clustering algorithms to generate multiple templates for each vocabulary item. While recognition accuracies obtained from multiple speakers compare favorably to those obtained from equivalent speaker-dependent systems, extension to speaker-independence is not foreseeable. Such an extension would require knowing when the training data were large enough so that they adequately account for all allowable pronunciations. Furthermore, assuming a sufficient data base could be collected, it is not clear that the recognition system will find, from amongst all of the acoustic variability present, all of the allophonic variants.

While the techniques mentioned are important engineering advances, they are not sufficient for extension of these systems to continuous speech recognition. That is, there is still no mechanism for dealing with feature assimilation effects between word boundaries. Presently, feature assimilation between phonemes is accounted for by choosing words as the recognition unit, possibly storing multiple or averaged templates for each word, and requiring sufficient silence (usually 200 msec) between words so that there is no feature spreading between them. To recognize continuous speech, template matching systems basically ignore feature spreading effects between words and use isolated word templates to spot words in the utterance (Myers and Rabiner, 1981). Although these systems have had limited success (greater than 94% string accuracy in a restricted digit task when the string length is known), this type of "brute force" approach cannot cope with some of the feature assimilation effects often seen at word boundaries (discussed in Section 1.1). Thus, extensions along these lines are unlikely.

In addition to this drawback, isolated word template-matching systems are unable to focus on phonetically relevant information needed to distinguish between acoustically similar words such as "way" and "lay," where the vowels in the word pair are the same and the consonants, although different, share common acoustic characteristics. This problem is the result of the distance metrics employed. Presently, in comparing two word templates, all parts of the utterance are weighted equally. Thus, in the example cited above, too much weight is given to the similarity in the frame-by-frame variations of the steady state vowel and too little weight to the differences between the consonants. As a result, for reasonable performance, the recognition vocabulary must consist of acoustically distinct words. This poses yet another problem for template-matching systems in that the size of the recognition vocabulary must be limited, since the acoustic distinctiveness between words decreases as the number of words increases.

During the past several years, many researcher have been investigating another approach for isolated word recognition systems which is based on hidden Markov models (HMM). With this technique, a labeled training data base is used to build Markov models for each word. In recognition, a probability score is computed for each word HMM given the unknown token, and the recognized word is the one whose model probability is highest. In a comparison of a speaker-independent isolated word recognition system based on HMM with one based on pattern-matching techniques with DTW, Rabiner et al. (1983) found that the HMM system performed slightly worse. It was hypothesized that this difference in performance was due to insufficient training data.

The most successful use of HMM to date has been in the speaker-dependent continuous speech recognition system developed at IBM (Jelinek et al., 1975; Jelinek, 1976; Jelinek, 1981). Recognition rates of 91% have been obtained for words selected from sentences in the 1000 word vocabulary of Laser Patent Text. Instead of word HMM models, this system uses HMM to model the time-varying spectral properties of phonetic segments. Each word in the lexicon is then represented as a sequence of phoneme models in a finite state graph, and feature assimilation between phonemes is handled through rules.

While consistently high word-recognition rates have been obtained with the IBM system for speakers who have trained the system extensively before use, extension of its approach to speaker-independence is problematic. Presently, the signal representation used to train the phone HMM consists of raw spectra which, as we said earlier, are intrinsically speaker dependent. Thus, to model all of the variability seen across all speakers would require an inordinate amount of training data and comparable computation and memory requirements.

1.2.3 Combined Methods

Over the past few years, efforts have been made to incorporate explicit speech knowledge into the mathematically-based frameworks. Below we discuss two such efforts which have reported positive results.

One effort which combined speech-specific knowledge and statistics is the FEATURE system developed by Cole et al. (1983). Instead of spectral templates, FEATURE used about 50 acoustic properties to recognize the isolated letters of the English alphabet. Motivated from a study of a large data base, these properties consisted of measures such as formant frequencies extracted from vowel regions and voice-onset time extracted from consonant regions. To integrate the properties for recognition, a statistical pattern classifier was used. For letters in the easily confused E set (B,C,D,E,G,P,T,V and Z), FEATURE obtained error rates of only 10% as compared to traditional spectral template matching systems which have error rates of between 30% and 40%.

A more recent system which combines these approaches was developed at BBN (Schwartz et al., 1985). In this speaker-dependent recognizer, context-dependent HMM models are used to recognize phonemes in continuous speech. However, in addition to the raw spectra, acoustic properties extracted from the speech signal are used within the HMM formalism to aid in certain phonetic contrasts. With this addition, confusions between acoustically similar phonemes decreased by as much as a factor of two. For example, Schwartz et al. state that the correct identification of the unvoiced stop consonants /p,t,k/ increased from 83% to 91%.

1.3 Thesis Scope

A conclusion which can be drawn from the literature review is that research in acoustic phonetics is of primary importance if speaker-independent continuous speech recognition systems shall be realized. More specifically, systematic studies of large data bases, combined with solid theoretical models of speech production and perception, are needed to uncover the proper acoustic properties for features and to gain an understanding of feature assimilation effects. Such a study is the focus of this research.

In this thesis we develop a framework for a phonetically-based speech recognition system. We view the recognition process as consisting of four steps. First, the features needed to recognize the speech sounds of interest must be specified. Second, the features must be translated into acoustic properties which can be quantified. Third, algorithms must be developed to automatically and reliably extract the acoustic properties. Finally, these properties must be combined for recognition.

The task we have chosen to study is the recognition of the semivowels /w,y,r,l/. This is a particularly challenging problem since the semivowels, which are acoustically very similar to the vowels, almost always occur adjacent to a vowel. As a consequence, spectral changes between these sounds are often quite gradual so that acoustic boundaries are usually not apparent. In this respect, recognition of the semivowels is more difficult than recognition of other consonants.

We have limited the recognition task to semivowels which are voiced and nonsyllabic. Devoiced allophones, which may occur when the semivowels are in clusters with unvoiced consonants, are excluded since some aspects of their acoustic manifestation are considerably different from that of the other semivowel allophones. In addition, the syllabic allophones of /r/ and /l/ in words like "bird" and "bottle" are excluded since they are more correctly classified as vowels.

To make this study manageable, we have simplified the semivowel recognition problem in several other ways. In particular, the recognizer is designed using polysyllabic words excised from the simple carrier phrase "________ pa." We chose this simple context as opposed to isolated words or more continuous speech because it allows for a more controlled environment. That is, following the test words with "pa" reduces the possibility of glottalization and other types of variability that occur in utterance-final position. In addition, since there is no sentence context to help convey the speaker's message, he or she is more likely to enunciate the words more clearly. Thus, the acoustic cues signalling phonetic contrasts should in general be more salient.

Although the recognition task has been simplified, it remains quite challenging. The data base chosen contains the semivowels in a variety of phonetic environments so that variability similar to that observed in continuous speech due to stress and feature assimilation is also found in the polysyllabic words. Thus, the methods used to recognize the semivowels are extendible to more continuous speech. This extension of the system is demonstrated with a small corpus of sentences.

The first part of this thesis lays the groundwork for the recognition system. We describe the data bases used to develop and test the recognition algorithms in Chapter 2. Also included in this chapter is a brief discussion of the tools used at different stages of this research.

Once a data base was collected to develop the recognition algorithm, we conducted an acoustic study to supplement data in the literature regarding the acoustic correlates for features needed to recognize the semivowels. The results of this study and a discussion of feature spreading and its apparent relation to syllable structure are given in Chapter 3.

After we identify acoustic properties for features, steps three and four of the framework outlined above are implemented. A description of how these steps are carried out is given in Chapter 4.

Chapter 5 contains an overview and a breakdown of the recognition results obtained for each of the data bases. The discussion therein points out the weaknesses and strengths of the recognition system. In addition, an analysis of the misclassifications brings forth several issues regarding feature spreading, attaching phonetic labels to patterns of features before lexical access, and using hand-transcribed data to evaluate recognition systems. The chapter closes with a comparison between the semivowel recognition results obtained in this thesis and those obtained in two earlier phonetically-based systems.

Finally, In Chapter 6, we summarize the results and discuss further some of the issues highlighted by this research. In particular, we discuss ideas regarding future studies of feature assimilation and lexical access from acoustic properties.

Chapter 2

Data Bases and Tools

This chapter describes the corpora used to develop and evaluate the semivowel recognition system. In addition, we discuss some of the tools used in various stages of this research. Among these tools is a formant tracker which we discuss in more detail since it was developed as a part of this thesis.

2.1 Data Bases

The initial step in this research was the design of a data base for developing and testing the recognition algorithms. Using ALEXIS, a software tool for lexicon search (Zue et al., 1986), we chose 233 polysyllabic words from the 20,000-word Merriam-Webster Pocket Dictionary. These words contain the semivowels and other similar sounds, such as the nasals and, in some contexts, other voiced consonants, in a variety of phonetic environments. They occur in word-initial and word-final positions such as the /y/ and /l/ in "yell," in intervocalic positions such as the /r/ and /l/ in "caloric," and adjacent to voiced (sonorant and nonsonorant) and unvoiced consonants such as the /w/ in the /dw/ cluster in "dwell," the /r/ and the /w/ in "carwash," the /y/ adjacent to the /n/ in "banyan" and the /r/ in the /str/ cluster in "astrology." In addition, the semivowels occur adjacent to vowels which are stressed and unstressed such as the word-initial /l/ and the prevocalic /l/ in "loathly," and they occur adjacent to vowels which are tense and lax, high and low, and front and back. An alphabetical listing and a grouping of the words according to various contexts are given in Appendix A. Some words occur in more than one of the categories based on context. The purpose of this overlap was to minimize the number of words in the data base while covering most contexts.

According to the phonetic transcription of the words given in the Pocket dictionary, the data base should contain 145 tokens of /r/, 139 tokens of /l/, 94 tokens of /w/ and 61 tokens of /y/. However, the actual number of semivowel tokens enunciated by each speaker differs because some words have multiple allowable pronunciations and some words were mispronounced. For example, a word which has a vowel-to-vowel transition where the first vowel has a /y/ offglide may be spoken with a /y/ inserted. Thus, the word "radiology" can be pronounced as [re^{y}di^{y}yɑləǰi^y] with an intervocalic /y/ or as [re^{y}di^{y}ɑləǰi^y] without an intervocalic /y/. Similarly, if the first vowel in a vowel-to-vowel transition has a /w/ offglide or is the retroflexed vowel /ɝ/, then a /w/ or an /r/ may be inserted, respectively. Thus, the word "flour" may be pronounced as [flɑ^{w}wɝ] with an intervocalic /w/ or as [flɑ^{w}ɝ] without a well enunciated /w/. Likewise, the word "laceration" may be pronounced as [læsɝre^{y}šən] with an /r/ inserted or as [læsɝe^{y}šən] without an /r/ inserted. In addition, a postvocalic /l/, when followed by another consonant, may not be clearly articulated. Thus, the word "almost" may be pronounced [ɔlmo^{w}st] or [ɔmo^{w}st]. Furthermore, a postvocalic /l/ which follows a reduced vowel may be articulated as a syllabic /l̩/. Thus, "unilateral" may be pronounced as [yunəlæɾəɾl̩] with a syllabic /l/, or it may be pronounced as [yunəlæɾəɾəl] with a postvocalic /l/. Finally, one of the speakers systematically confused /r/ and /w/. For example, the intervocalic /w/ in "rauwolfia" was replaced by an /r/ and the prevocalic /r/ in "requiem" was replaced by a /w/.

For these reasons, judgement regarding the inclusion or exclusion of a semivowel is often ambiguous. Several measures were used to make this decision if a semivowel was not clearly heard when the utterance or a portion thereof was played. First, within the region in question, we looked for significant formant movement towards values expected of the semivowel. Second, we looked for other spectral changes such as a decrease in energy since the semivowels are usually weaker than adjacent vowels. Finally, we sometimes consulted with other transcribers.

For acoustic analysis and the development of the recognition algorithms, each word was recorded by two males (one black and one white) and two females (one black and one white). The speakers are from the northeast (New York and Rhode Island) and the midwest (Ohio and Minnesota). They were recorded in a quiet room with a pressure-gradient close-talking noise-cancelling microphone. The microphone was placed about 2 cm in front of the mouth at a right angle just above the midline. All of the words were hand-transcribed to facilitate the acoustic study (see Chapter 3). When transcribing the data base, we placed markers at particular instants of time to divide the speech signal into segments which were assigned labels that in some way described some property(s) of the delineated regions.

Two corpora were used to test the recognition system. The first data base consisted of the same polysyllabic words spoken by two additional speakers (one female, one male, both white) from the same geographical areas cited above. The same recording set-up was used. These words were also transcribed to facilitate the evaluation of the recognition algorithms. The second data base consisted of a small subset of the sentences contained in the TIMIT data base (Lamel et al., 1986). In particular, we chose the sentences "She had your dark suit in greasy wash water all year" (Sent-1) and "Don't ask me to carry an oily rag like that" (Sent-2), since they contain several semivowels in a number of contexts. Presently, the TIMIT data base is being segmented and labelled by several experienced transcribers with the help of an automatic alignment system (Leung and Zue, 1984). From the transcribed utterances, we selected 14 repetitions of Sent-1 (6 females and 8 males) and 15 repetitions of Sent-2 (7 females and 8 males). The speakers cover 7 U.S. geographical areas and an "other" category used to classify talkers who moved around often during their childhood. Like the words in the other data bases, these sentences were recorded using a close-talking microphone.

2.2 Tools

The semivowel recognition system was implemented on the MIT Speech Communication Group's LISP machine facility for which several software tools have been developed to aid speech research. The way in which the tools were used in this thesis is described briefly in this section. A more detailed discussion of the tools is offered in (Zue et al., 1986).

2.2.1 SPIRE

Initial processing of the data base was done with the Speech and Phonetics Interactive Research Environment (SPIRE). First, the recorded words were digitized using a 6.4 kHz low pass filter and a 16 kHz sampling rate. Such a wide frequency range helps in the identification of obstruents (stops, fricatives and affricates) and, therefore, in

Table 2.1: Symbols Available in SPIRE for Phonetic Transcription

Unvoiced Stops:	p t k č
Voiced Stops:	b d g ǰ
Stop Gaps:	p° t° k° ʔ b° d° g° ɾ
Nasals:	n m ŋ r̃
Syllabic Nasals:	n̩ m̩ ŋ̩ l̩
Unvoiced Fricatives:	s š f θ
Voiced Fricatives:	z ž v ð
Glides:	l r w y
Vowels:	iʸ ɪ ɛ eʸ æ ɑ ɑʷ ɑʸ ʌ ɔ ɔʸ oʷ ʊ u ü ɝ
Schwa:	ə ə̥ ɨ ɚ
H, Silences:	h ɦ ▯ □
Special Markings:	# * $ + - ' " ~

the discrimination between sonorant and nonsonorant sounds. This is approximately the frequency range that is often used for spectrogram reading. Second, the speech signals were preemphasized to compensate for the relatively weak spectral energy at high frequencies, particularly for sonorants. This preemphasis means that the average spectral energy is similar at the higher and lower frequencies. Finally, SPIRE was used to transcribe the data bases. The set of symbols available for phonetic transcription is shown in Table 2.1. Most of these symbols were taken from the International Phonetic Alphabet (IPA). However, there are some additions and modifications. For example, the word initial sound in "yell" is denoted by the symbol /y/ in SPIRE and by the symbol /j/ in the IPA. In addition, the syllabic /l/ as in "table" is denoted by the symbol /l̩/ in SPIRE and by /ɫ/ in the IPA.

Although there are 58 phonetic symbols in Table 2.1, we found this list incomplete for some of the feature-spreading phenomena occurring between semivowels and adjacent segments. These effects are described below.

- The features of a vowel or consonant and a following /r/ may overlap considerably, such that the acoustic manifestation of these two segments is an r-colored vowel or an r-colored consonant, respectively. An example of this phenomenon is shown in Figure 2.1, which compares spectrograms of the words "harlequin" and "marlin" spoken by the same person. In the case of "marlin," the lowest frequency of F3 clearly occurs in the /r/ region which follows the vowel. However, in "harlequin," F3 is lowest at the beginning of the vowel and remains steady for a considerable duration of the vowel, after which it rises due to the influence of the /l/. In the latter case, an /r/ segment separate from the vowel segment is not apparent. Thus, instead of forcing nonoverlapping

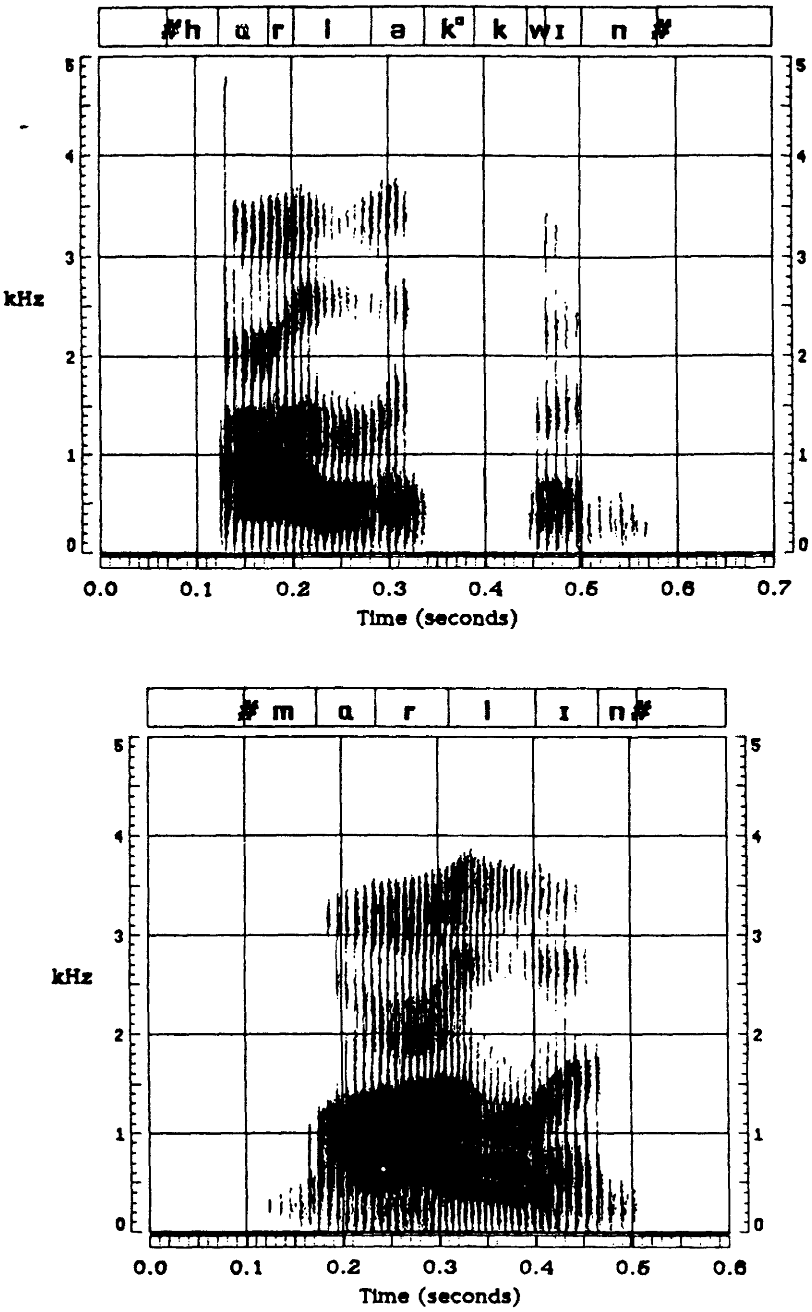


Figure 2.1: A comparison of the words "harlequin" and "marlin." In "harlequin" the underlying /ɑ/ and /r/ sounds appear to be merged into one segment, in the sense that the lowest point of F3 occurs at the beginning of the vowel. Thus, the transcription should allow overlapping sounds. In "marlin," F3 is well above 2000 Hz in the beginning of the /ɑ/, and it falls steadily to its lowest point in the /r/. Thus, the /ɑ/ and /r/ appear to be separate segments.

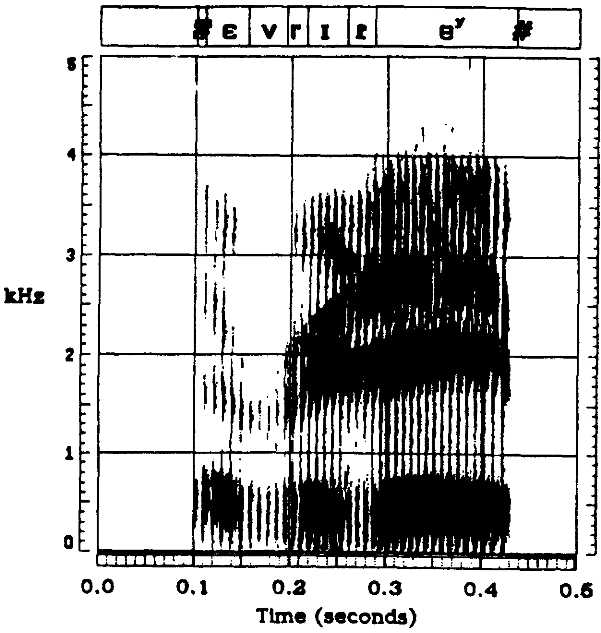


Figure 2.2: The /v/ in "everyday" appears to be sonorant and retroflexed. In fact, the lowest point of F3 occurs within this segment. Thus, the /v/ and /r/ appear to overlap.

juxtaposed segments, a more correct transcription facility would allow the transcribed /ɑ/ and /r/ regions to be combined into one region with an appropriate r-colored vowel label. A similar example of this phenomenon, in this case the retroflexed consonant /v/, is illustrated in Figure 2.2, where a spectrogram of the word "everyday" is shown.

- When in a cluster with unvoiced consonants, the semivowels are sometimes devoiced. An example of this type of feature spreading is shown in Figure 2.3, which compares spectrograms of the word "queen" spoken by two different speakers. In the spectrogram on the top, the /w/ in the /kw/ cluster is only partially devoiced such that there are considerable F2 and F3 transitions from the /w/ into the following vowel. However, in the spectrogram on the bottom, the /w/ is completely devoiced. In this case, little in the way of F2 and F3 transitions occur between the fricated /w/ and the following vowel. Instead, the acoustic cues indicating the presence of the /w/ are the low-frequency frication and the low-frequency burst of the /k/. As in the case described above, these phonetic segments co-occur, causing segmentation to be difficult.

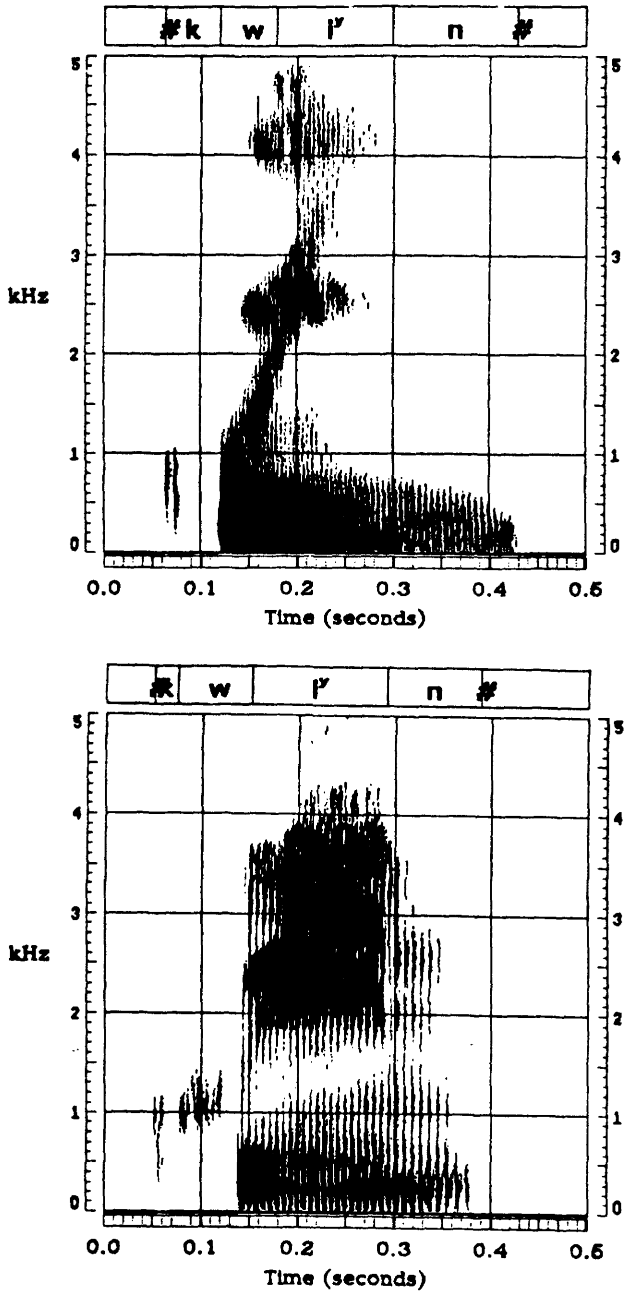


Figure 2.3: Two spectrograms of the word "queen," spoken by different speakers. In the example on the top, the /w/ is only partially devoiced. In the example on the bottom, the /w/ is completely devoiced.

Since SPIRE does not have phonetic symbols for devoiced semivowel allophones that occur simultaneously with unvoiced consonants, and since the convention within the speech research group regarding this phenomenon is to label some portion of the beginning of the vowel region as being the devoiced semivowel, part of the vowel was transcribed as /w/. To locate the beginning of the fricated /w/, we successively removed frames from the beginning of the word until the /k/ was no longer audible, so that we heard /wi^{y}n/.

2.2.2 SEARCH

SEARCH (Structured Environment for Assimilating the Regularities in speeCH) is an exploratory data analysis tool which facilitates use of several statistical techniques for examination of a large body of data. For example, questions such as "What percentage of the intervocalic semivowels have significantly less energy than their adjacent vowels?" can be answered with this tool. In acoustic analysis, this software package was used in several ways. First, we used it to study the effectiveness of parameters in capturing properties observable in spectrograms. Second, SEARCH was used to determine the relationship between an acoustic property and the context of a particular phonetic segment or class of phonetic segments. Finally, since SEARCH can display data in various forms including histograms, scatter plots and a bar-like display, we used it to determine thresholds for quantifying the extracted properties.

2.2.3 Knowledge-Based Formant Tracker

Although it is not yet offered as a general tool, a formant tracker implemented in the SPIRE facility was developed as a part of the thesis. We based the formant tracker on peak-picking of the second difference of the log-magnitude linear-prediction (ISDLM-LP) spectra. Since the development of this software turned out to be a major undertaking, a discussion of the strategy, techniques and constraints employed in the automatic formant tracker is given below.

Strategy

Since we are interested in the recognition of voiced and sonorant semivowels, formant tracking is performed only in those regions specified by the voiced and sonorant detectors (for the parameters used, see Section 3.2.3). To obtain initial estimates of the formant frequencies, a strategy similar to that developed by McCandless (1974) is used. A block diagram of this strategy is given in Figure 2.4.

Before formant tracking, energy peaks, which usually correspond to syllabic nuclei within vowel regions, and energy dips, which usually correspond to syllable boundaries within sonorant consonant regions, are detected (a discussion of how they are obtained is given in the subsection "Intersonorant Semivowels" of Section 4.3.1). Peak picking begins at an energy peak since the formants are most likely to be tracked correctly in the middle of a vowel region, which is least affected by feature assimilation effects such as nasalization or retroflexion. First, the algorithm back tracks, filling formant slots with peaks based on continuity constraints (the frame rate is one per 5 msec) until a boundary is reached. In this case, a boundary can be either the beginning of the detected voiced sonorant region or an energy dip. Second, the algorithm forward tracks from this energy peak, deciding on peaks in each successive frame until a boundary is reached. In this case, a boundary can be either the end of the detected voiced sonorant region or an energy dip. If there are other energy peaks within the voiced sonorant region, this process is continued until the formants have been tracked in each frame.

Techniques

As mentioned above, we chose to pick peaks from the ISDLM-LP spectra. We decided to use this spectral representation of the vocal tract transfer function for several reasons. First, the semivowels are articulated orally with no side branches (except possibly for /l/). Thus, in the frequency range of interest, the transfer function of these sounds can be represented accurately by an all-pole model. Second, spurious peaks which are common in many forms of spectral analysis are rare in the linear prediction spectra, and, therefore, they are rare in the ISDLM-LP spectra. Thus, peak-picking is a more tractable problem using a linear-prediction-based spectra. Finally, shoulder resonances, which occur often in linear prediction spectra and usually cannot be detected through peak picking, show up as distinct peaks in the ISDLM-LP spectra (Christensen et al., 1976).

Although this spectral representation reduces the peak merger problem, this problem as well as problems due to nasalization still remain. In the former case, two peaks which are completely merged in the linear prediction spectra will also be completely merged in the ISDLM-LP spectra. As a result, there will be empty formant slots. In such instances, we compute the ISDLM-LP spectra inside the unit circle. An iterative

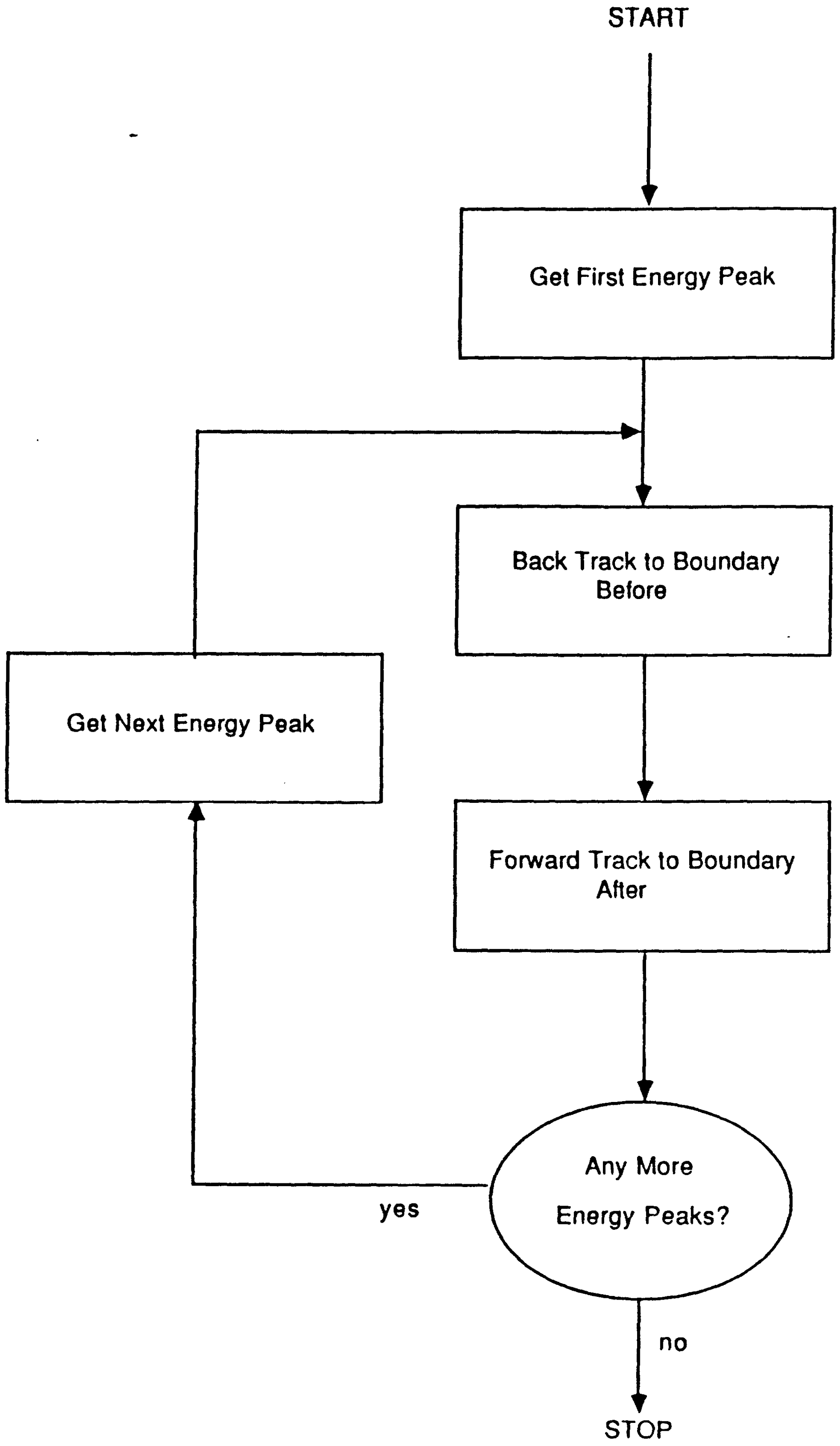


Figure 2.4: Block diagram of formant tracking strategy within a voiced sonorant region.

procedure is used to resolve the merged peaks. The enhanced spectrum is first computed with a radius of 0.996. If the missing peak has not been resolved, the radius is decremented by 0.004 and a new enhanced spectrum is computed. This process is continued until either the missing peak has been resolved or the radius is less than 0.88. Most merged peaks will be resolved through this type of enhancement. However, in a few instances, further enhancement may be needed to resolve a missing formant. In addition, in some cases, LPC may represent very close peaks by one pole pair. A higher order LPC model is needed to resolve such peaks.

This missing-formant problem also occurs in nasal consonants. However, in this case, the missing formant is not due to merged peaks; it is missed because the formant has been cancelled by a zero.

Whether they are due to our inability to resolve them or zero cancellation, these missing formant slots are filled in through interpolation in the final steps of the formant tracker. This process is discussed below.

Constraints

Both frequency and amplitude constraints are used to decide which peaks to identify as formants. Before formant tracking, we estimate the pitch frequency of the speaker to determine whether the talker is male or female. The pitch detector (which is part of the SPIRE facility) was developed by Gold and Rabiner (1969). Based on this pitch frequency estimate, we use empirically-determined male or female formant anchors for F1, F2, F3 and F4. These anchors are used to decide on the peaks in the frames marked by the energy peaks. When back tracking or forward tracking from this frame, continuity constraints as well as frequency thresholds, which restrict how much a formant can change within 5 msec, are used to decide which peaks will go into the formant slots. Due to continuity constraints and using the strategy outlined in Figure 2.4, the decision of which peaks are assigned to the formant slots in the frame marked by the energy peak(s) is crucial. An incorrect decision in this frame will result in unreasonable formant tracks. To minimize the chances of making a wrong decision in this frame, we compute the ISDLM-LP spectra on the unit circle and at several radii inside the unit circle. In most cases, this procedure guarantees that all merged formants are resolved.

Amplitude constraints are used when two peaks are competing for the same formant slot. This situation happens when a nasal formant is piesent within a vowel region or, in the case of females, when there is a strong harmonic below F1. In most instances, the amplitude of the nasal formant or the harmonic is weak compared to the amplitude of the adjacent peak(s). Thus, in each frame, we always choose the strongest peak to go into the formant slot being considered, unless it does not meet the continuity constraints.

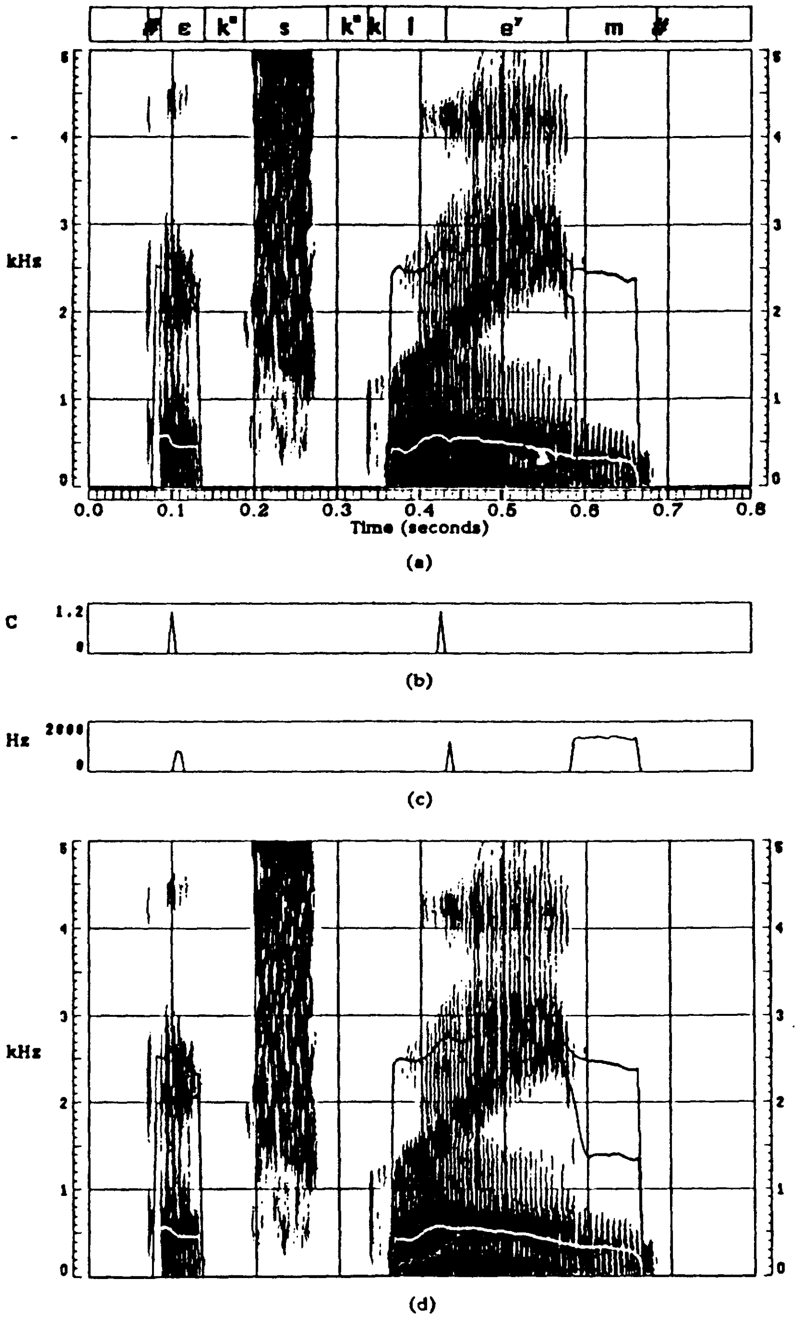
Even with enhancement, and frequency and amplitude constraints, incorrect decisions are sometimes made. Once the formants have been tracked throughout the voiced sonorant regions within the utterance being analyzed, the formants are processed by an algorithm which tries to ensure reasonable formant tracks. This algorithm is described in the next section.

Post-Processor

When formant tracking, the first five peaks in the ISDLM-LP spectra are candidates for formant slot positions. Four of the five peaks are assigned to slots for F1, F2, F3 and F4. The peak not assigned to any of these positions is not thrown away, but is kept in either a slot labelled "possible nasal formant" or a slot labelled "extra peak." If the frequency of the additional peak is less than the frequency of the peak assigned to the F2 slot, then it is placed in the possible nasal slot. Otherwise, it is placed in the extra slot. Thus, the extra slot usually contains F5, and the possible nasal slot may contain either a nasal formant (or the peak it was competing with, usually F1), a spurious low frequency peak, or, in the case of females, a strong harmonic.

These extra peak slots are used in the post-processor. In this stage of processing, the formant tracks are checked for discontinuities. If one or more tracks possess a discontinuity, and if substitution or partial substitution of the tracks in either of the extra peak slots will result in a more continuous track, they are switched. If such a switch occurs between any one of the formant tracks and either of the extra peak slots, then each formant track is checked again for discontinuities. This process is continued until no change occurs for any of the formant tracks.

Two situations in which this post processing stage was necessary to obtain reasonable formant tracks are illustrated in Figures 2.5 and 2.6 for the words "exclaim" and "plurality," respectively. In both cases, the outputs of the formant tracker, and the formant tracker plus the post processing stage and smoothing are compared. Also shown in the figures are the locations of the energy peaks and energy dips used to compute the formant tracks and the extra peaks obtained. For the word "exclaim,"



(a)

(b)

(c)

(d)

Figure 2.5: An illustration of the performance of the post processing stage in the tracking of the word "exclaim." (a) Formant tracks obtained before post processing and smoothing. Note that F2 is zero within the /m/. (b) Location of energy peaks used in formant tracker. (c) Spectral peaks occurring in the "possible nasal formant slot." (d) Formant tracks after post processing and smoothing. Note that the peaks occurring in the "possible nasal formant slot" between 600 msec and 680 msec have been placed in the empty F2 slots.

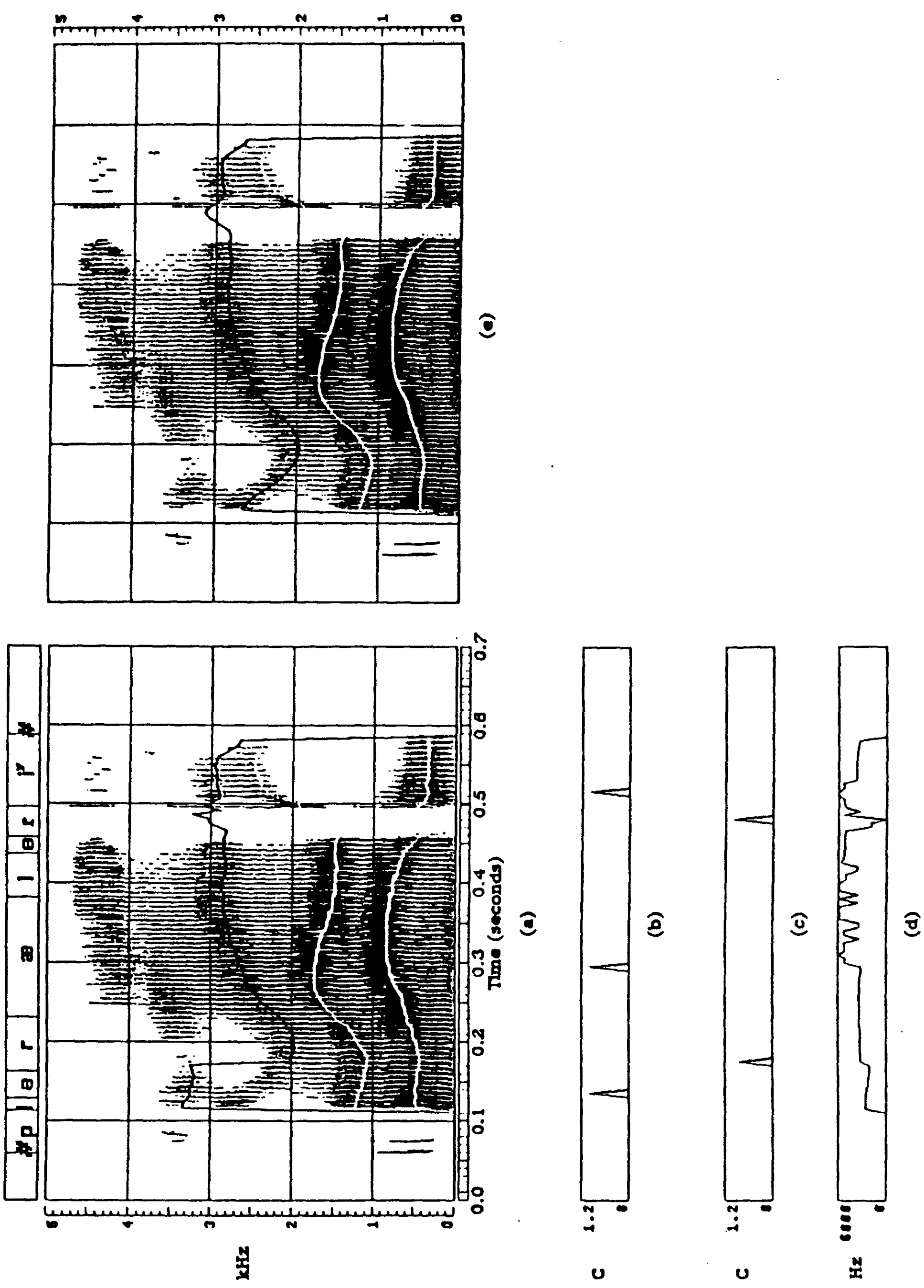


Figure 2.6: An illustration of the performance of the post processing stage in the tracking of the word "plurality." (a) Formant tracks obtained before post processing and smoothing. (b) Location of energy peaks used in formant tracker. (c) Location of energy dips used in formant tracker. (d) Spectral peaks occurring in the extra formant slot. (e) Formant tracks obtained after post processing and smoothing. Note that the peaks occurring in the extra formant slot between about 110 msec and 170 msec were placed in the F3 track.

no peaks are stored in the F2 slot during the nasal sound /m/ before the post processing stage. Instead, due to the large discontinuity in F2 (a change of about 900 Hz) between the vowel /e^y/ and the /m/, this information is stored in the extra slot for possible nasal formants. However, after post processing, this information is placed in the F2 slot.

In the case of the word "plurality" which was spoken by a female speaker, F3 and F4 (2500 Hz and 3250 Hz, respectively) at the time of the first energy peak are both close to the anchor frequency for F3 (2930 Hz). Since F4 is about 4 dB greater in amplitude, it was placed in the formant slot for F3, and F3 was placed in the extra formant slot. As can be seen, this resulted in a sharp discontinuity at 170 msec within the F3 track. However, during one iteration of the post processor, the peaks placed in the F3 slot before the discontinuity were replaced by the information stored in the extra peak slot. From part d, we see that the corrected F3 track is always in the F3 range observable from the spectrogram.

Interpolation and Smoothing

Even with enhancement, the problem of peak mergers and the additional problem of nasalization result in frames with missing formants. After the post-processing stage (discussed above), the tracks obtained for F1, F2 and F3 are checked for missing data. If any of these tracks have missing data, a polynomial is used to fit the the formant track in a region surrounding the frames with missing data. This region is defined by formant tracks on each side of the missing data where the sign (positive or negative) of the slope is constant for several frames. The order of the least mean-square polynomial used to fit the data depends upon the sign of the slopes of the tracks on both sides of the missing data. If the slopes on both sides are postive or negative, then linear interpolation is done. However, if the slopes differ in sign, a second order polynomial is used for interpolation.

Once the missing data have been filled in through interpolation, the formant tracks of F1, F2 and F3 are smoothed twice with the zero phase filter

$$F_i'(n) = \tfrac{1}{4}F_i(n-1) + \tfrac{1}{2}F_i(n) + \tfrac{1}{4}F_i(n+1).$$

Two situations in which interpolation was needed are shown in Figures 2.7 and 2.8 which contain formant tracks for the words "harlequin" and "urethra." For several frames in the word "harlequin," F3, because it has a low amplitude, could not be

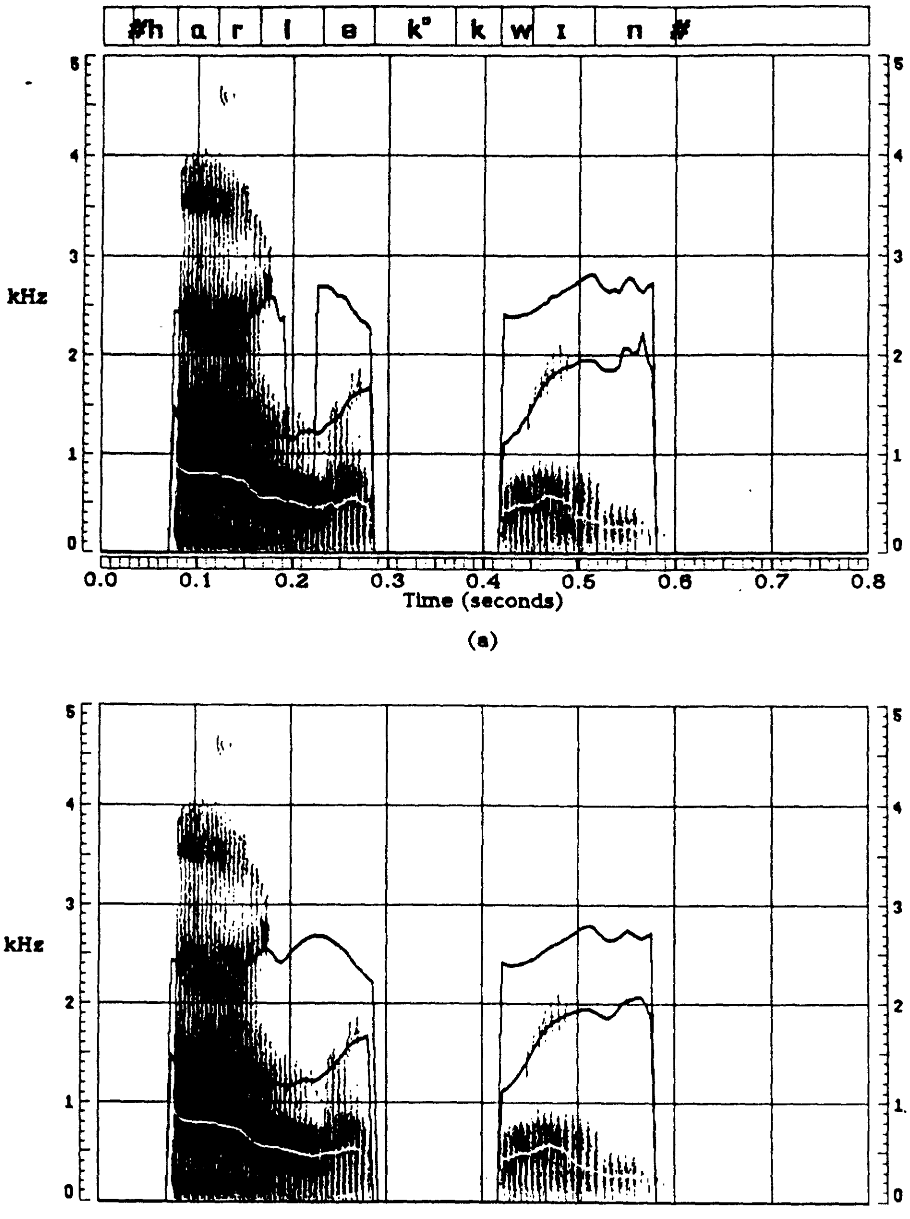


Figure 2.7: An illustration of the performance of the interpolation algorithm. (a) Formant tracks obtained for "harlequin" before interpolation. Note that F3 during the /l/ was not tracked. (b) Formant tracks for "harlequin" after interpolation and smoothing.

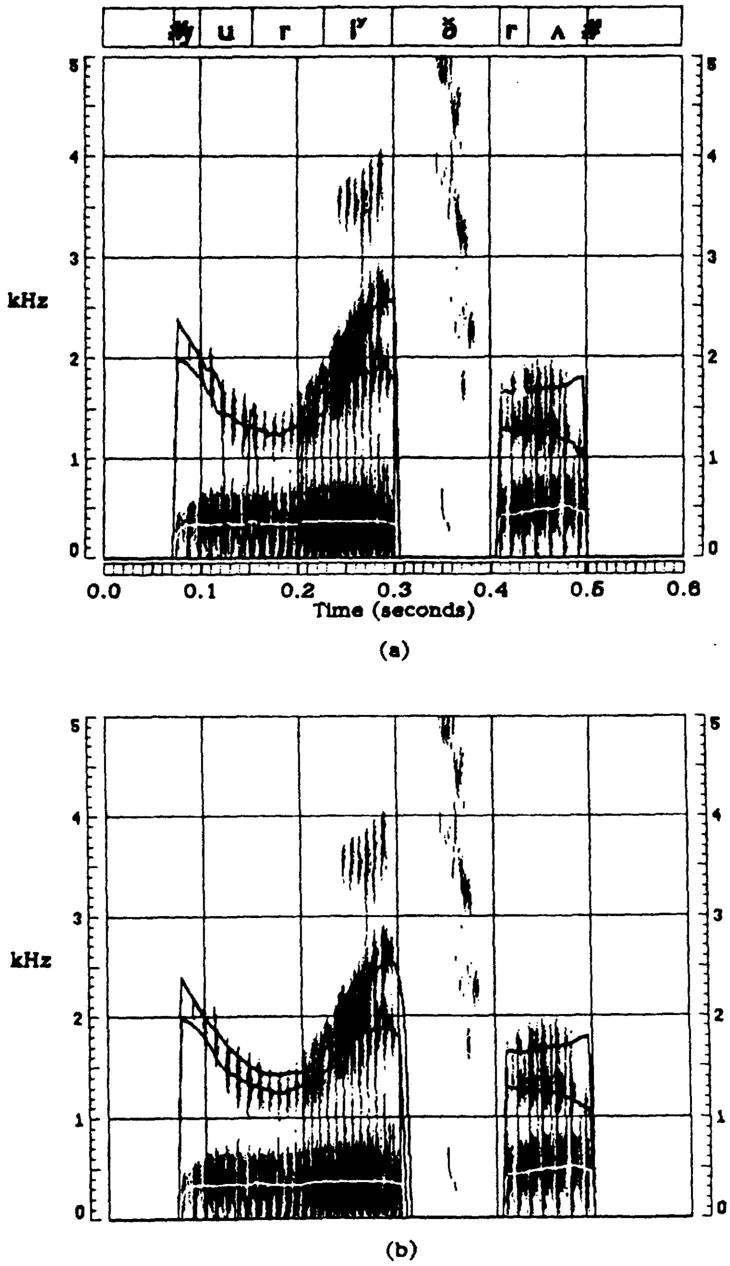


Figure 2.8: An illustration of the performance of the interpolation algorithm. (a) Formant tracks obtained for "urethra" before interpolation. Note that F3 during the /u/ and /r/ segments was not always tracked. (b) Formant tracks for "urethra" after interpolation and smoothing.

tracked, even with enhancement. However, by using the frequency values of F3 on both sides of the missing frames, reasonable estimates of F3 were obtained through interpolation. Likewise, for the word "urethra," F3 was not tracked for several frames during the /u/ and /r/. In this case, however, F2 and F3 were merged in the LPC spectra such that enhancement did not resolve F3. Again, reasonable estimates of F3 were obtained through interpolation.

Performance

To refine the formant tracker, incorrect tracks obtained for the words said by a particular speaker were corrected by modifying the code. Errors were detected by overlaying the tracks on a spectrogram and by comparing the formant estimates with the peaks occurring in wide-band and narrow-band short-time spectra. After reasonable formant tracks were obtained for all words, F1, F2 and F3 were computed for the words said by a different speaker. Again, errors were corrected by refining the code. This process continued until reasonable tracks were obtained across all of the words said by all of the speakers of the database used to develop the recognition algorithms.

For the other two corpora, estimates of the formant tracks were computed once. We have not looked at all of the formant tracks to determine the number of errors that occurred. However, the results obtained in different stages of the recognition process (discussed in Chapters 4 and 5) have led us to the discovery of formant-tracking errors occurring within semivowels. In the corpus containing polysyllabic words, incorrect tracks were obtained for 1.4% of the 850 semivowels. In addition, 10 words were not tracked at all due to a minor problem which has since been corrected. In the corpus of sentences, incorrect tracks were obtained in 1.4% of the 141 semivowels. In this case, one sentence was not tracked at all.

Chapter 3

Properties of Semivowels

3.1 Introduction

The sounds /w,y,r,l/ are called semivowels because they have properties which are similar to both vowels and consonants. Like the vowels, the semivowels are produced orally without complete closure of the vocal tract and without any frication noise. Furthermore, the rate of change of the formants and of other aspects of the spectrum tends to be slower than that of the other consonants and the degree of constriction needed to produce these sounds does not inhibit spontaneous voicing. Thus, as in the case of vowels, a voiced steady state (with a duration that is usually in the range 30 msec to 70 msec) is often observed from spectrograms of the semivowels. These acoustic properties can be observed in Figures 3.1 and 3.2 (Zue, 1985) where, along with x-ray tracings of the vocal tract, we show spectrograms of these sounds in word-initial position within the two sets of minimal pair words "we," "ye," "reed" and "lee" and "woo," "you," "rue" and "Lou."

The semivowels /l/ and /r/ are often referred to as liquids; their articulation involves contact of the blade and/or tip of the tongue with the alveolar ridge. In the production of /l/, a lateral constriction is made by placing the center of the tongue tip against the alveolar ridge. In addition, when they occur before vowels, there is usually a rapid release of the tongue tip from the roof of the mouth. As a result, an abrupt spectral change between /l/'s and following vowels is often observable from a spectrogram (Fant, 1960; Dalston, 1975). This phenomenon can be seen at the boundary between the /l/ and the following vowels in Figure 3.2.

In the production of /r/, the constriction is made toward the back of the alveolar

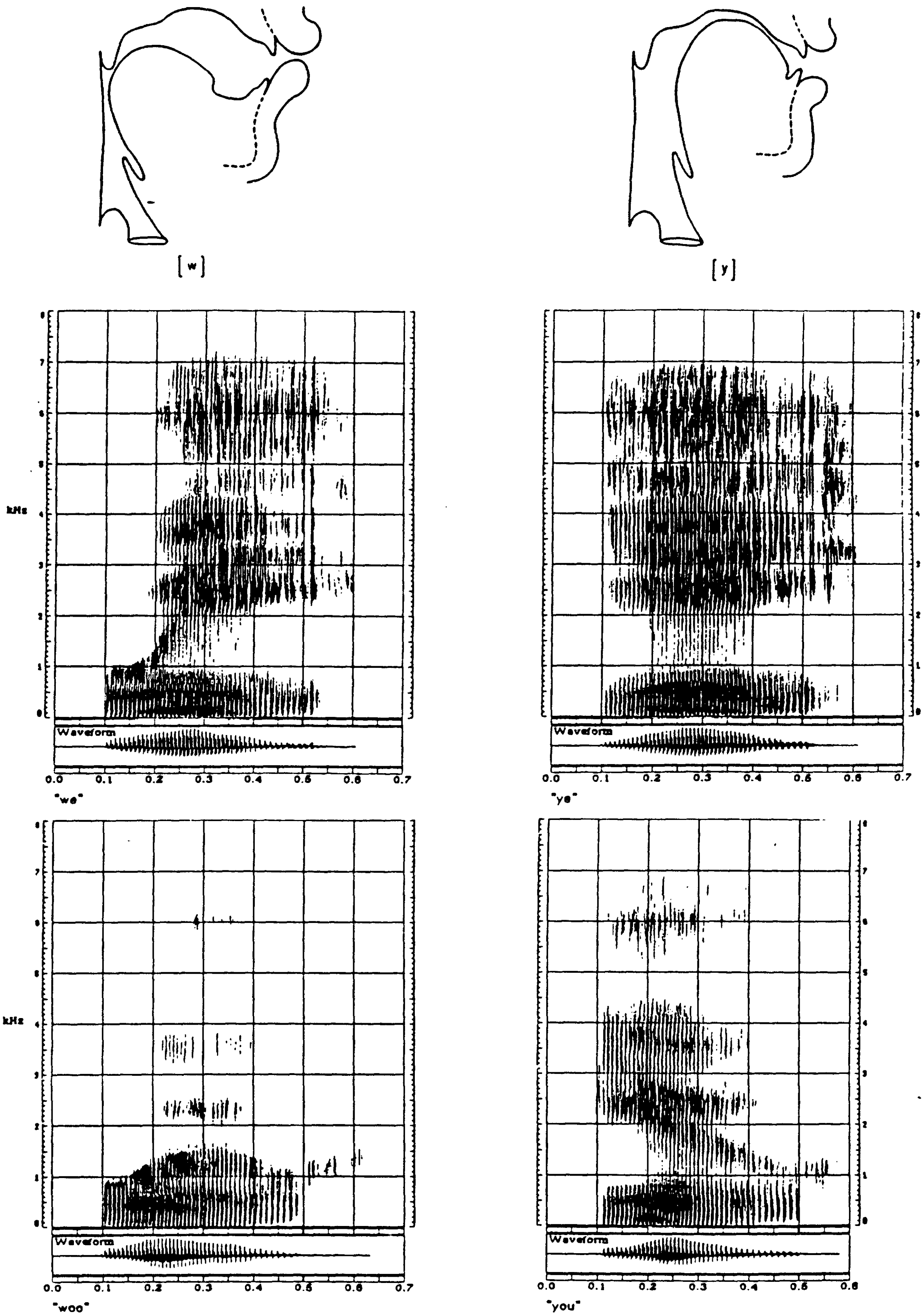


Figure 3.1: X-ray tracings of the vocal tract and wide band spectrograms of the words "we" and "ye" (top), and "woo" and "you"(bottom).

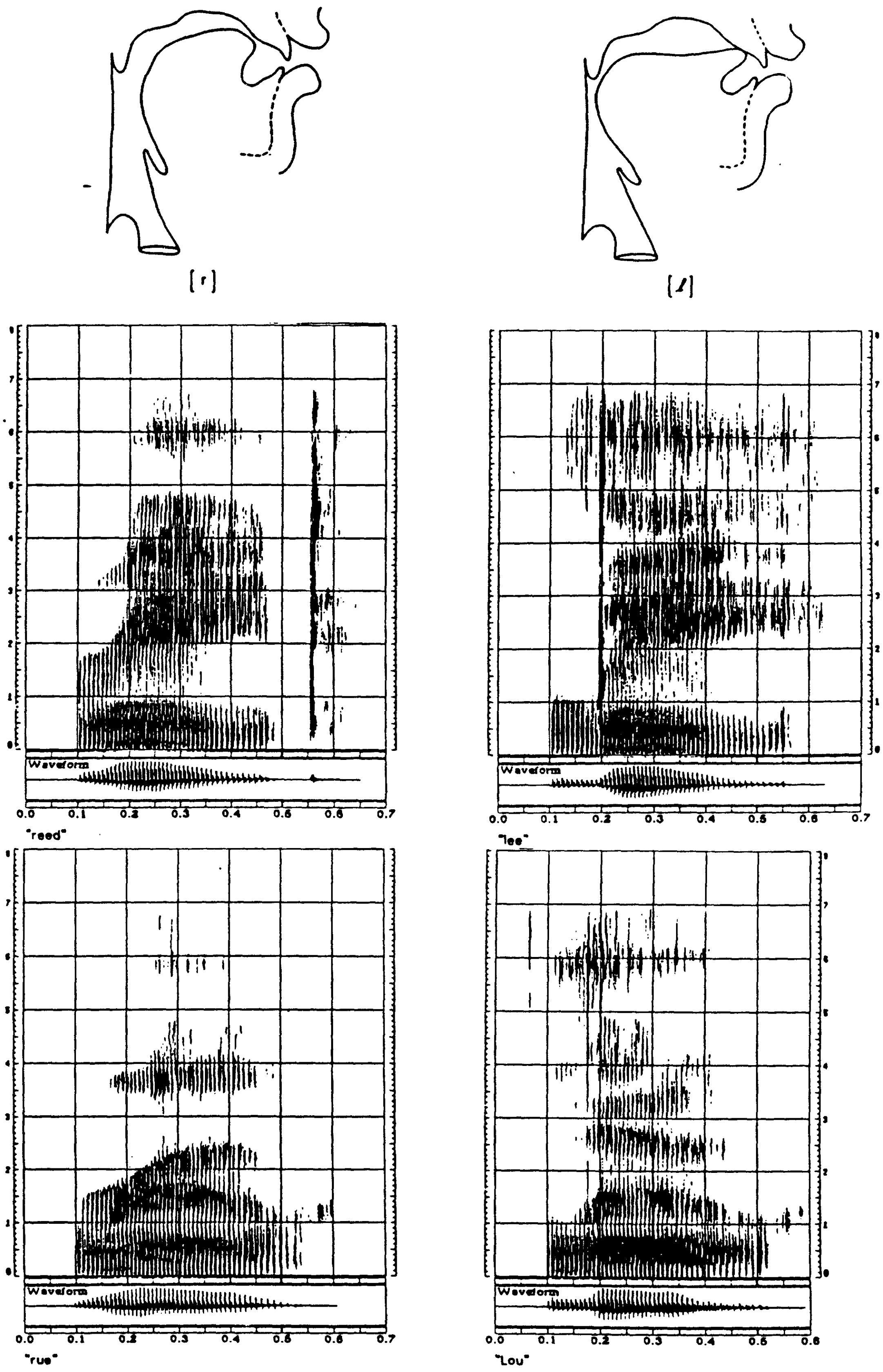


Figure 3.2: X-ray tracings of the vocal tract and wide band spectrograms of the words "reed" and "lee" (top), and "rue" and "Lou" (bottom).

ridge, near the palate. This placement of the tongue tip creates a sublingual cavity whose lowest natural resonance is usually at or below 2000 Hz and close to the lowest natural resonance of the back cavity (Stevens, in preparation). These two resonances constitute F2 and F3. This acoustic distinctiveness of /r/ can be seen in Figure 3.2.

The semivowels /w/ and /y/ are produced with a vocal tract configuration similar to those for the vowels /u/ and /i/, respectively, but with a more radical constriction. As a result, /w/ has lower F1 and F2 frequencies than /u/, and /y/ has a lower F1 frequency and a higher F2 frequency than /i^y/. These differences can be seen in the words "woo" and "ye" of Figure 3.1.

The semivowels /w/ and /y/ are often referred to as glides or transitional sounds because they are produced as the articulators move towards or away from an articulation. That is, they are considered as onglides when they precede vowels (i.e., the /y/ in the word "compute") or offglides when they follow vowels (e.g., the second component of the diphthong /ɔ^y/ in the word "boy"). In addition, the glides are often intermediate sounds when the articulators pass from the position of one vowel, with the appropriate offglide, to the position of another vowel. An example of this is the /y/ sound often heard in the pronunciation of "the ice," due to the /y/ offglide of the vowel /i^y/. The glides are produced with constant movement of the articulators such that the formants in the transition between them and adjacent vowels exhibit a smooth gliding movement accompanied by either an increase in amplitude when they occur before vowels, or a decrease in amplitude when they are the offglides of vowels. The semivowel /r/ is sometimes included in the definition of a glide. However, /l/ is usually not included since, as mentioned above, the spectral change between a prevocalic /l/ allophone and the following vowel is usually abrupt.

In addition to exhibiting a difference in manner of articulation, the semivowels differ from other consonants from a distributional standpoint as well. The semivowels must occupy a position in a syllable immediately adjacent to the vowel, with the exception of words like "snarl" in which the /r/ occurs between the vowel and the word-final /l/. (Some acoustic data obtained in the study suggest that there should not be such an exception clause in the phonotactic constraints of semivowels. For further discussion, see Section 3.3.) Furthermore, the semivowels are the only consonants that can be the third member of a three-consonant syllable-initial cluster.

Like the other consonants, however, the semivowels usually occur at syllable margins. That is, they generally do not have or constitute a peak of sonority (sonority, in this case, is equated with some measure of acoustic energy). The relatively low amplitude of the semivowels as compared to the vowels is due in part to the fact that they tend to have a low frequency first formant. It may also be due to a large F1 bandwidth caused by the narrower constriction, or to an interaction between the vocal folds and the constriction (Bickley and Stevens, 1987). At present, this phenomenon is not well understood.

3.2 Acoustic Study

There have been many acoustic and perceptual studies involving some or all of the semivowels (Lisker, 1957; O'Connor et al., 1957; Lehiste, 1962; Kameny, 1974; Dalston, 1975; Bladon and Al-Bamerni, 1976; Bond, 1976). Mainly, these studies have focused on the acoustic and perceptual cues which distinguish among the semivowels and the coarticulatory effects between semivowels and adjacent vowels. We have used the acoustic and perceptual findings of this past work to guide an acoustic study of the semivowels and other sounds contained in the data base (see Chapter 2) designed for the thesis.

In this study, we attempt to quantify some of the findings of past acoustic and perceptual research using energy based parameters, formant tracks and fundamental frequency. While the parameters were selected on the basis of some informal work, we realize that there may be other ones which better capture the desired acoustic properties.

Most of the measurements made in the study are relative. That is, a measure either examines an attribute in one speech frame in relation to another frame, or, within a given frame, examines one part of the spectrum in relation to another. As a result, the relative measures tend to be independent of speaker, speaking rate and speaking level.

The following sections are organized by measure(s). First, we discuss measures which help to distinguish between the semivowels. These measures are based on formant frequencies and formant transitions. Second, we discuss measures which help to separate the semivowels from other classes of sounds. These measures are based on bandlimited energies and measures of the rate of spectral change.

The features for which the measures are presumed to be correlates of are mentioned in each section. However, a summary of this study is given in Chapter 4 in Table 4.3.

This table includes the features needed to separate the semivowels as a class from other sounds and to distinguish between the semivowels, the acoustic properties for features, and the parameters from which these relative measure are extracted.

To conduct the study of the semivowels, we used the tool SEARCH (see Section 2.2.2). Recall that this tool is token-based such that the measurements are dependent upon the hand transcription of the words.

3.2.1 Formant Frequencies

Past studies agree that important cues for distinguishing among the semivowels are the frequencies of the first three formants (F1, F2 and F3). Given minimal pair words, F1 separates the glides /w/ and /y/ from the liquids /l/ and /r/, F2 separates /w/ from /l,r/ from /y/, and F3 separates the liquids /l/ and /r/. The data in this study concur with these observations. The formant frequencies were estimated by averaging samples around the time of a minimum or maximum in a formant track within the hand-transcribed semivowelregion. In the case of /w/ and /l/, the values of F1, F2 and F3 were averaged around the time of the minimum value of F2. For /y/, the formant values were averaged around the time of the maximum value of F2. Finally, for /r/, the formant values were averaged around the time of the F3 minimum. At most, three samples were used to compute the average. The results are shown for each speaker and across speakers in Tables 3.1-3.5 for word-initial, prevocalic (including semivowels that are word-initial and adjacent to a voiced consonant), intervocalic, postvocalic (including the /l/ in words like "snarl") and word-final semivowels. Speakers SS and SM are females and speakers MR and NL are males.

Also included in Tables 3.1-3.5 are the normalized formant values (F1-F0, F2-F1, F3-F0 and F3-F2) which are used in the recognition system discussed in Chapter 4. In addition, the distributions of the normalized formants are shown in Figures 3.3, 3.4 and 3.5 for the prevocalic, intervocalic and postvocalic semivowels, respectively. The formants were normalized in this manner to better capture some of the acoustic properties of the semivowels. The acoustic correlates of the features *back* and *front* are usually thought of in terms of the spacing between F1 and F2, rather than the absolute frequency of F2. Similarly, results from preliminary work suggest that, in addition to the frequency of F3, the spacing between F3 and F2 is important in establishing the acoustic correlate of the feature *retroflex*. We observed that /w/'s can have F3 values comparable to that of some /r/'s. However, F3 and F2 tend to be much closer for

Table 3.1: Average formant frequencies of word-initial semivowels broken down by speaker and averaged across all speakers.

	w	l	r	y
F1	365	443	389	308
F2	696	1250	1270	2040
F3	2170	2480	1620	2710

speaker: MR

	w	l	r	y
F1	374	393	345	294
F2	768	1100	1090	1960
F3	2340	2540	1490	2930

speaker: NL

	w	l	r	y
F1	319	397	340	287
F2	819	1420	1290	2350
F3	2420	2810	1880	3000

speaker: SM

	w	l	r	y
F1	324	384	360	240
F2	674	1110	969	2350
F3	2440	2730	1500	3480

speaker: SS

	w	l	r	y
F1	347	404	358	281
F2	739	1220	1150	2190
F3	2330	2640	1620	3040

all speakers

	w	l	r	y
F1 − F0	211	266	216	138
F2 − F1	392	821	794	1910
F3 − F0	2200	2510	1480	2900
F3 − F2	1600	1420	471	855

all speakers

Table 3.2: Average formant frequencies of voiced prevocalic semivowels broken down by speaker and averaged across all speakers.

	w	l	r	y
F1	347	423	401	301
F2	691	1030	1240	2040
F3	2160	2410	1630	2750

speaker: MR

	w	l	r	y
F1	381	394	370	323
F2	788	1060	1150	2010
F3	2320	2510	1590	2780

speaker: NL

	w	l	r	y
F1	339	387	366	311
F2	782	1200	1360	2330
F3	2440	2850	1970	2970

speaker: SM

	w	l	r	y
F1	337	386	392	266
F2	697	1060	1120	2350
F3	2370	2600	1650	3100

speaker: SS

	w	l	r	y
F1	351	397	383	305
F2	793	1090	1220	2190
F3	2320	2600	1710	2910

all speakers

	w	l	r	y
F1 − F0	214	258	242	163
F2 − F1	388	693	835	1890
F3 − F0	2180	2460	1570	2770
F3 − F2	1580	1510	491	719

all speakers

Table 3.3: Average formant frequencies of intervocalic semivowels broken down by speaker and averaged across all speakers.

	w	l	r	y
F1	314	424	444	333
F2	652	934	1210	2110
F3	2230	2400	1570	2730

speaker: MR

	w	l	r	y
F1	383	445	441	326
F2	884	1050	1200	2010
F3	2270	2580	1670	2750

speaker: NL

	w	l	r	y
F1	344	441	466	357
F2	603	1140	1330	2490
F3	2470	2900	1950	3100

speaker: SM

	w	l	r	y
F1	350	466	482	389
F2	718	1090	1220	2360
F3	2370	2670	1650	3010

speaker: SS

	w	l	r	y
F1	349	445	460	361
F2	771	1060	1240	2270
F3	2340	2640	1720	2920

all speakers

	w	l	r	y
F1 – F0	211	305	317	213
F2 – F1	422	610	783	1910
F3 – F0	2200	2500	1570	2770
F3 – F2	1570	1580	473	648

all speakers

Table 3.4: Averaged formant values for postvocalic liquids broken down by speaker and averaged across all speakers.

	l	r
F1	454	487
F2	821	1240
F3	2380	1690

speaker: MR

	l	r
F1	459	486
F2	875	1280
F3	2690	1770

speaker: NL

	l	r
F1	493	528
F2	994	1350
F3	2830	2040

speaker: SM

	l	r
F1	457	509
F2	901	1330
F3	2620	1840

speaker: SS

	l	r
F1	465	503
F2	898	1300
F3	2630	1830

all speakers

	l	r
F1 − F0	323	363
F2 − F1	433	799
F3 − F0	2490	1690
F3 − F2	1740	531

all speakers

Table 3.5: Average formant values for word-final liquids broken down by speaker and averaged across all speakers.

	l	r
F1	444	484
F2	768	1270
F3	2430	1670

speaker: MR

	l	r
F1	454	444
F2	841	1240
F3	2680	1670

speaker: NL

	l	r
F1	481	472
F2	932	1350
F3	2830	2050

speaker: SM

	l	r
F1	443	484
F2	864	1330
F3	2590	1760

speaker: SS

	l	r
F1	455	471
F2	850	1300
F3	2630	1790

all speakers

	l	r
F1 – F0	313	330
F2 – F1	396	828
F3 – F0	2490	1650
F3 – F2	1780	493

all speakers

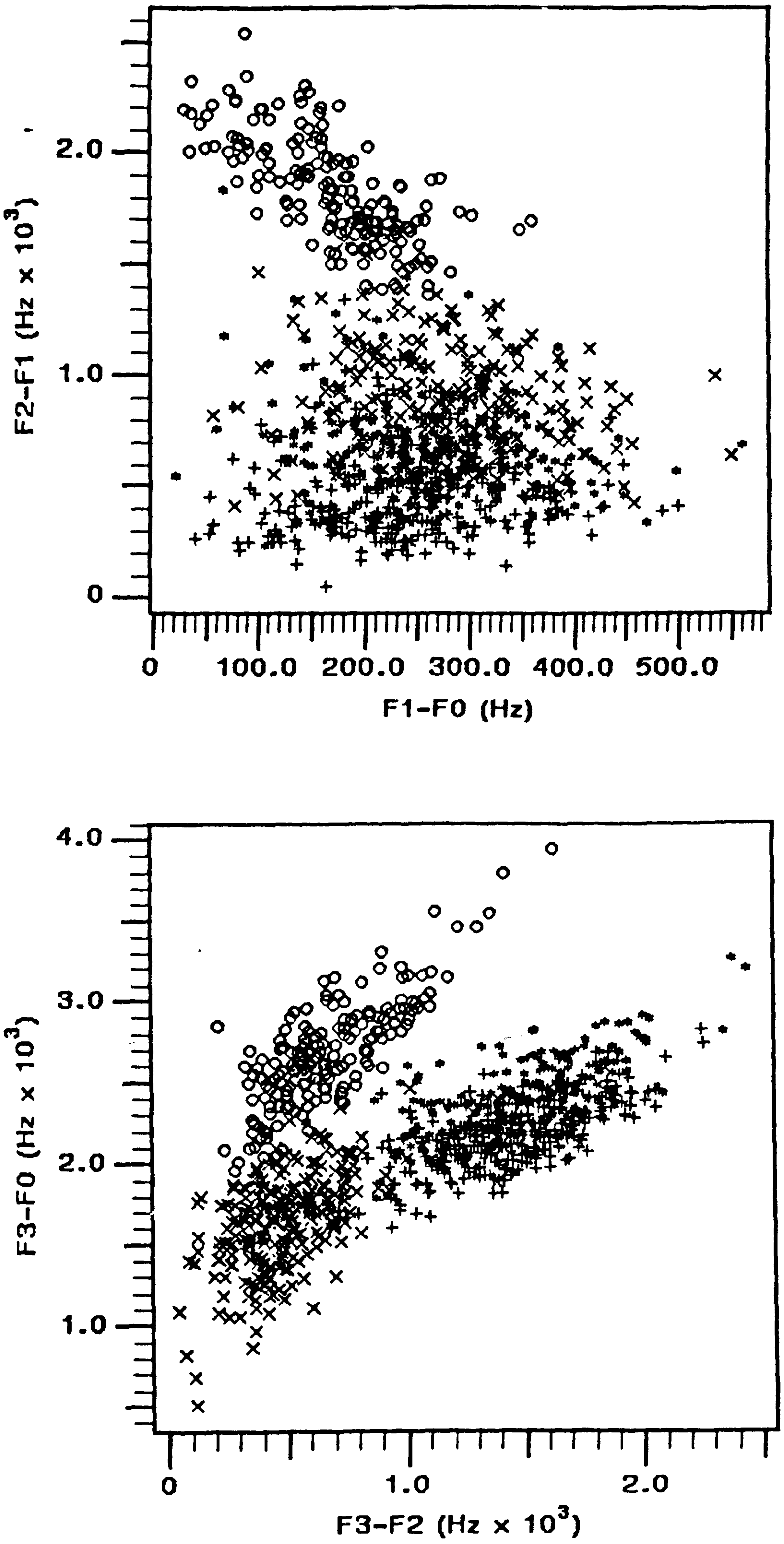


Figure 3.3: Plots of normalized formant values for prevocalic semivowels. w: +, y: ∘, r: ×, l: ∗.

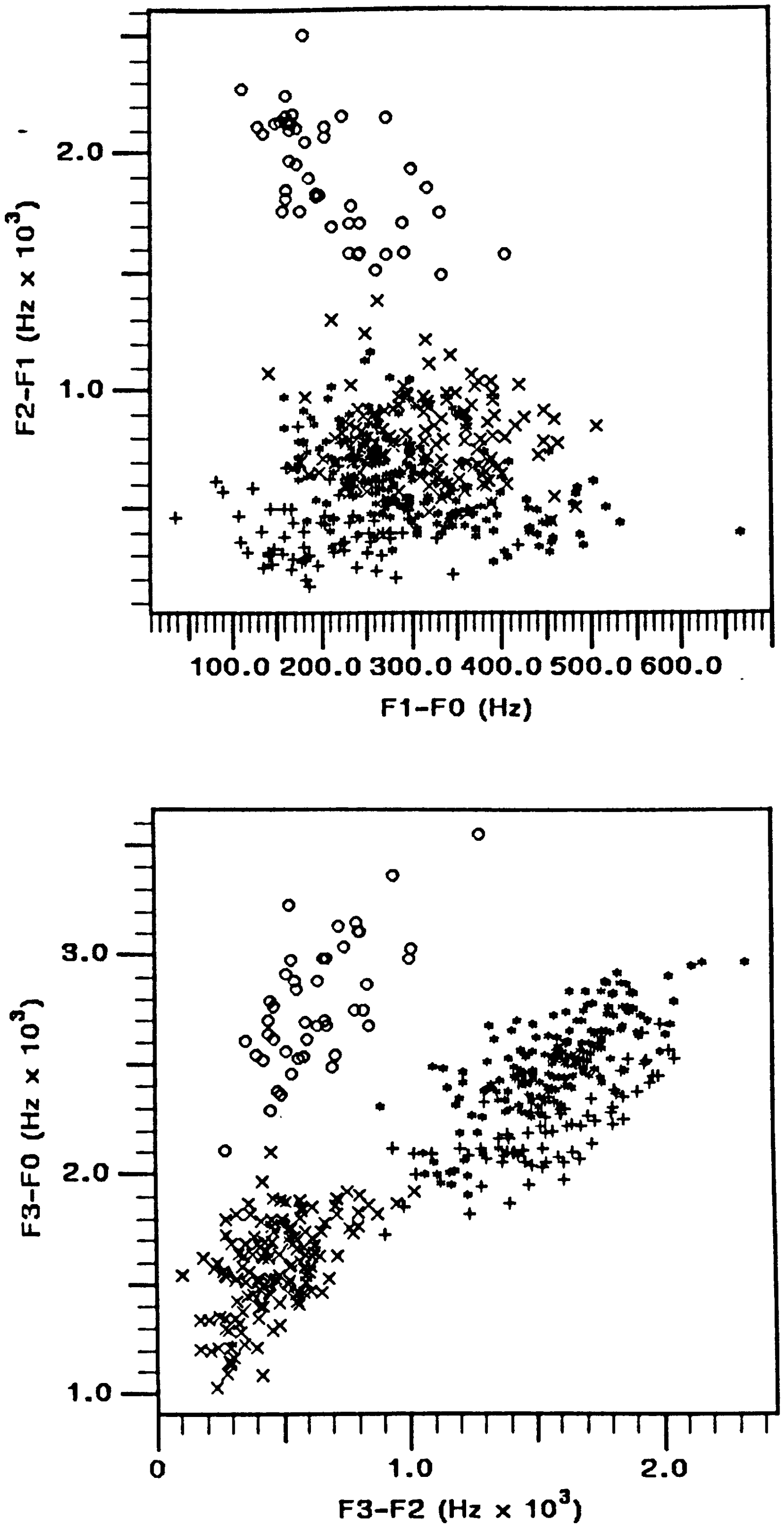


Figure 3.4: Plots of normalized formant values for intervocalic semivowels. w: +, y: o, r: ×, l: *.

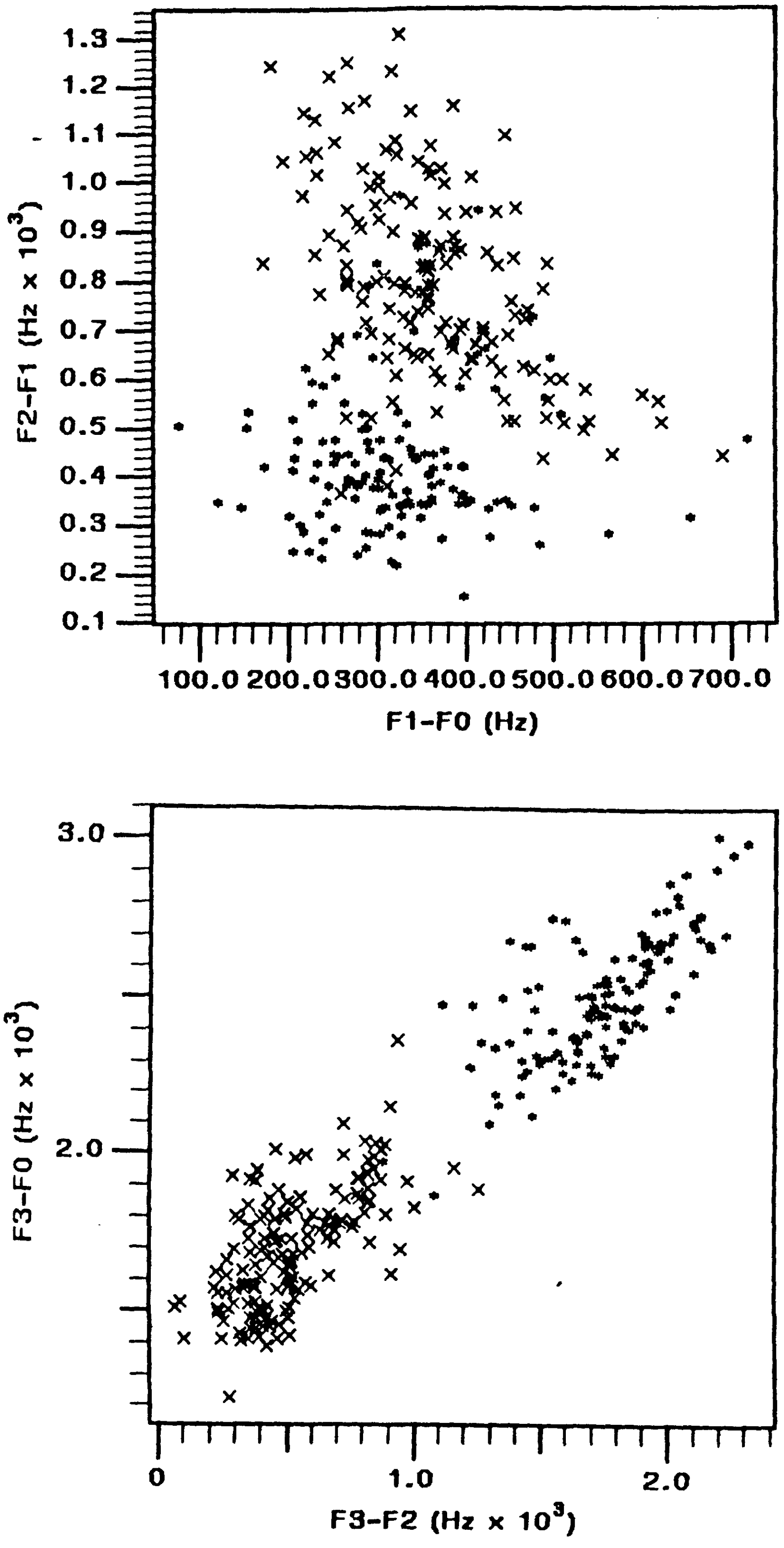


Figure 3.5: Plots of normalized formant values for postvocalic semivowels. w: +, y: o, r: ×, l: $*$.

/r/, whereas F2 and F1 tend to be much closer for /w/ (this difference between the acoustic properties of these sounds can be seen in the formant plots of Figures 3.3 - 3.5). Therefore, we included the measure F3-F2. In addition, while the acoustic correlates of the features *high*, *low* and *retroflex* relate to the frequencies of F1 and F3, the sex of the speaker is usually considered before making any judgements regarding their presence or absence. That is, since F1 and F3 will generally be higher for a female than for a male, we usually normalize for sex. In a simple attempt to account for the sex of the speaker, we normalized F1 and F3 by the average fundamental frequency, F0, computed across the voiced regions of the utterance. More specifically, we subtracted F0 from F1 and F3.

Several observations can be made from these data. First, the average formant frequency values of the word-initial, prevocalic and intervocalic semivowels are comparable. The generally higher F1 frequency for the intervocalic semivowels suggests that they are not usually as constricted as their prevocalic allophones. Second, the difference in the formant values for postvocalic and prevocalic /l/ and /r/ allophones support previous findings. That is, a postvocalic /l/ is more velarized than a prevocalic /l/, resulting in a much lower F2, a higher F1 and, therefore, a smaller F2-F1 difference. This allophonic variation is shown in Figure 3.6 where the word-initial /l/ in "loathly" is compared with the word-final /l/ in "squall." Both words were spoken by the same speaker. In the former case, the /l/ has F1, F2 and F3 frequencies of about 370 Hz, 990 Hz and 2840 Hz, respectively. In the latter case, the frequencies of F1, F2 and F3 are about 465 Hz, 700 Hz and 2660 Hz, respectively.

As for /r/, Lehiste found that the postvocalic /r/ allophone (all word-final with the exception of the /r/ in "wharf") has higher frequencies for F1, F2 and F3 than the word-initial /r/ allophone. Furthermore, Lehiste found that the average word-final F2 frequency for a postvocalic /r/ is in the range of F3 for a word-initial /r/ allophone, and that the average postvocalic F3 frequency is about 300 Hz greater than its average F2 frequency. Our data agree with most of these findings. F1, F2 and F3 of the postvocalic or word-final /r/ allophones are generally higher than their corresponding values for prevocalic or word-initial /r/ allophones, respectively. However, for speaker MR, the frequency values for F2 and F3 are similar for the word-initial and word-final /r/ allophones, and for the prevocalic and postvocalic /r/ allophones. This is also true for speaker SM if we compare F2 and F3 of the prevocalic and postvocalic /r/ allophones; however, these frequency differences are greater for the word-initial

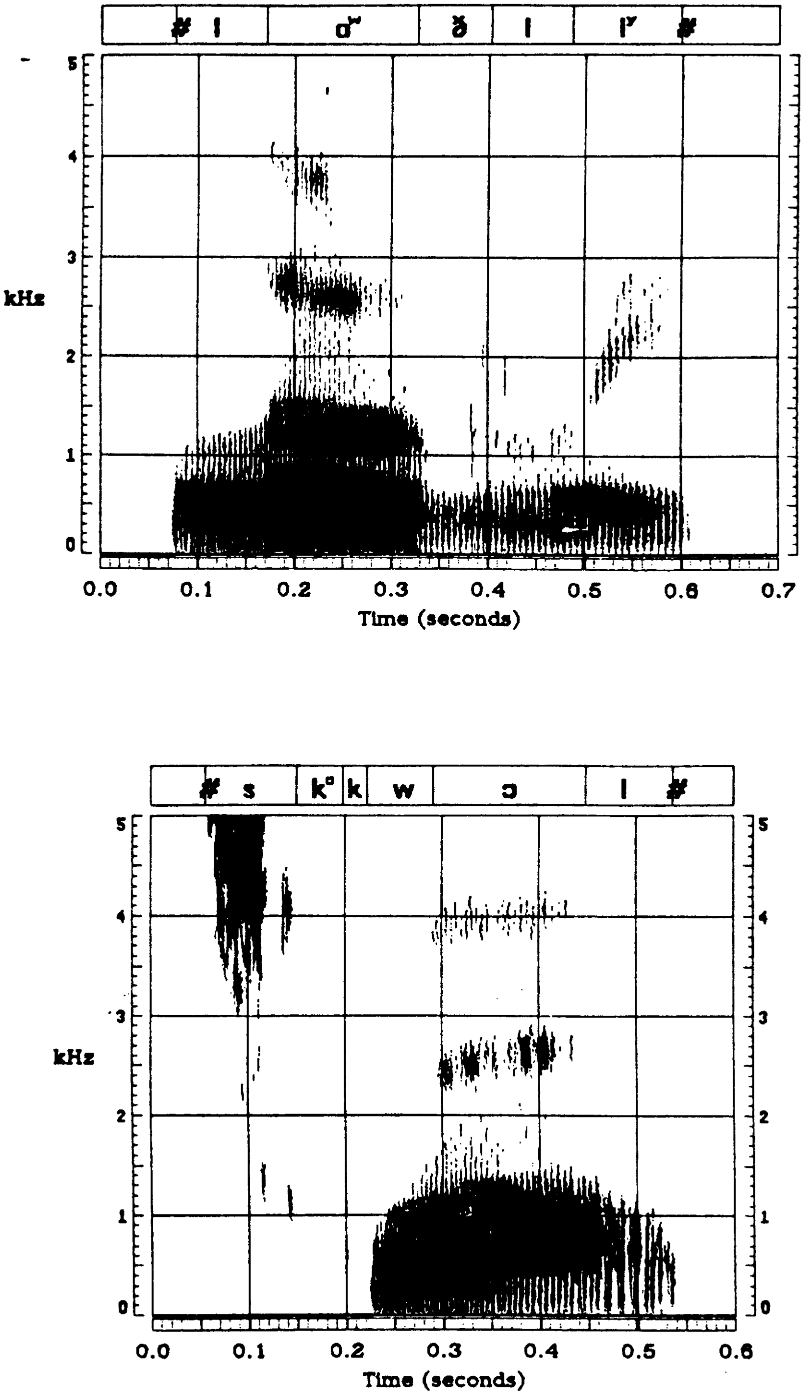


Figure 3.6: Wide band spectrogram of the words "loathly" and "squall."

and word-final /r/ allophones. Thus, comparing the F2 and F3 values obtained by averaging across all speakers the word-final and word-initial /r/ allophones, or the postvocalic and prevocalic /r/ allophones, we see that, unlike Lehiste's data, F2 of the /r/ allophone following a vowel is not close to F3 of the /r/ allophone preceding a vowel. Furthermore, the difference between F3 and F2 of the /r/ allophones which follow a vowel is about 500 Hz.

This allophonic variation can be seen in Figure 3.7 by comparing the formant frequencies of the word-initial /r/ in "rule" with the word-final /r/ in "explore." Both words were spoken by the same speaker. In the former case, the /r/ has F1, F2 and F3 frequencies of about 340 Hz, 1100 Hz and 1550 Hz, respectively. In the latter case, the word-final /r/ has F1, F2 and F3 frequencies of about 460 Hz, 1280 Hz and 1950 Hz, respectively.

Finally, the wide spread in the distribution of average formant values given in Figures 3.1, 3.2 and 3.3 for the prevocalic, intervocalic and postvocalic semivowels shows that the formant frequencies of the semivowels are affected by those of adjacent sounds. That is, the F1 frequency of the semivowels is usually lower than the average frequency of F1 when they are adjacent to high vowels, and usually higher than the average F1 value when they are adjacent to low vowels. Similarly, the F2 frequency of the semivowels tends to be lower than the average F2 frequency when they are adjacent to back vowels, and higher than the average F2 frequency when they are adjacent to front vowels. Furthermore, the F3 frequency of the semivowels /w/, /y/ and /l/ tends to be lower than their average value when they are either adjacent to /r/, such as the /w/ in "carwash," or they are one segment removed from an /r/, such as the /y/ in "Eurasian" and the /l/ in "brilliant." In addition, F3 of /r/ tends to be higher than its average value when it is adjacent to a front vowel(s). These contextual effects account for most of the overlap between /r/ and the other semivowels on the basis of F3-F0.

3.2.2 Formant Transitions

Given the average formant frequencies of the semivowels, certain formant transitions can be expected between them and adjacent vowels. To determine the direction and extent of this formant movement, the average semivowel formant values were subtracted from the average formant values of the adjacent vowel(s). The average vowel formant values were computed from the values occurring at the time of the maximum value of F1 within the hand-transcribed vowel region and the frequencies occurring

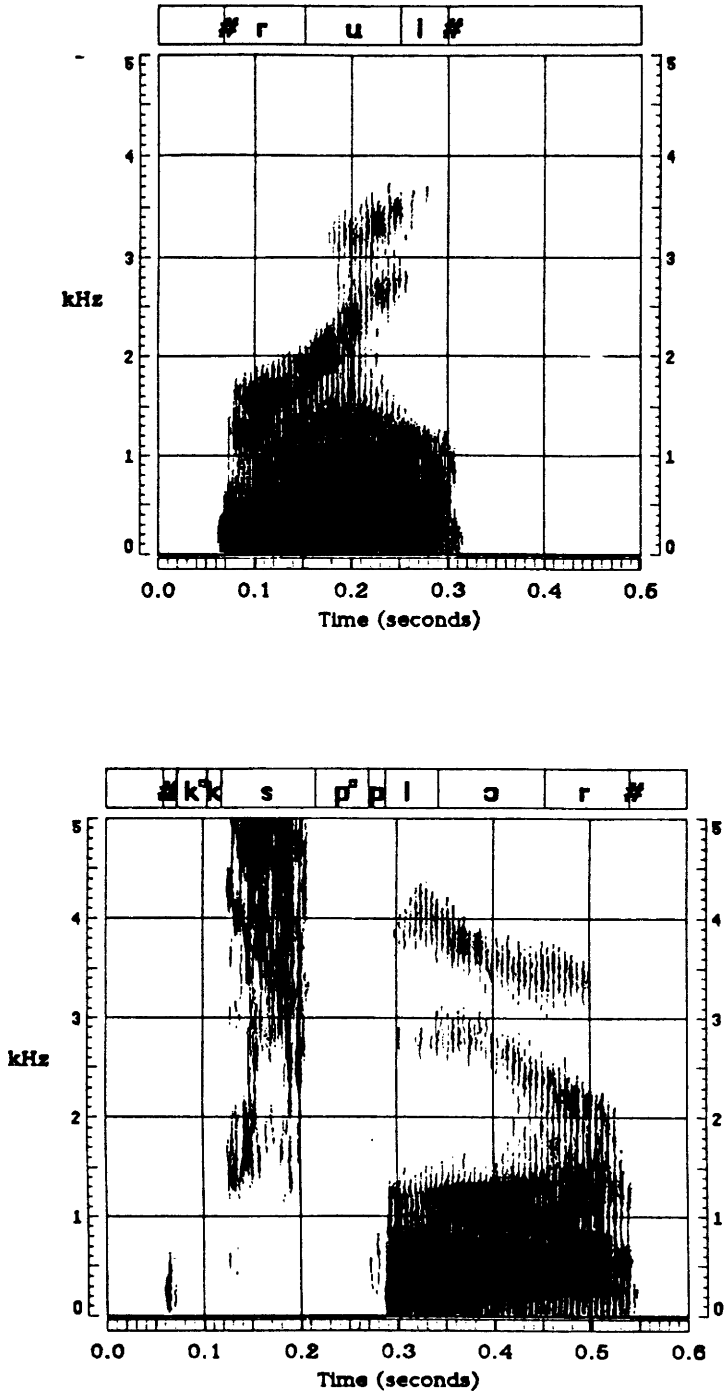


Figure 3.7: Wide band spectrogram of the words "rule" and "explore."

Table 3.6: Averages and standard deviations of the differences between the average formant values of prevocalic semivowels and those of following vowels.

	ΔF1		ΔF2		ΔF3	
	avg	std	avg	std	avg	std
w	194	124	516	275	17	315
y	175	135	-519	333	-503	393
l	158	123	436	308	-7	224
r	128	107	281	307	466	382

in the previous and following frames (if they also occur within the hand-transcribed region). The findings of this part of the acoustic study are shown in Tables 3.6, 3.7 and 3.8 for the average differences between the formant values of the vowels and adjacent prevocalic, intervocalic and postvocalic semivowels, respectively. Also included are the standard deviations. Below we discuss the results separately for each semivowel.

/w/

As expected, compared to the adjacent vowel, F1 and F2 are almost always lower for a /w/. However, the data for F3 show that the transition of F3 between a /w/ and an adjacent vowel can be positive or negative. A negative F3 transition from a /w/ into an adjacent vowel may seem surprising, since /w/ is produced labially. However, we found this to be the case mainly when /w/ is adjacent to a retroflexed vowel. The average change in F3 between prevocalic /w/'s and following retroflexed vowels is about -215 Hz. In the case of intervocalic /w/'s, the average increase in F3 from a preceding retroflexed vowel is about 300 Hz, and the average decrease in F3 into a following retroflexed vowel is about 200 Hz. Examples of this phenomenon can be seen in the spectrograms and formant tracks of the words "thwart" and "froward" which are displayed in Figure 3.8. Although F3, due to its low amplitude, is not always visible within the /w/, the direction of the F3 movement can be inferred from the visible transitions in the adjacent vowel(s), and it is apparent in the accompanying formant tracks.

Table 3.7: Average and standard deviation of the difference between the average formant values of intervocalic semivowels and those of the surrounding vowels.

preceding vowel

	ΔF1		ΔF2		ΔF3	
	avg	std	avg	std	avg	std
w	123	76	657	342	-36	326
y	108	130	-527	390	-499	400
l	103	93	314	237	-140	217
r	48	70	167	218	438	292

following vowel

	ΔF1		ΔF2		ΔF3	
	avg	std	avg	std	avg	std
w	169	134	619	303	-24	291
y	176	153	-524	334	-346	295
l	84	117	378	205	-8	136
r	57	110	264	274	433	324

Table 3.8: Averages and standard deviations of the differences between the average formant values of postvocalic liquids and those of the preceding vowels.

	ΔF1		ΔF2		ΔF3	
	avg	std	avg	std	avg	std
l	128	112	352	225	-159	217
r	68	90	39	269	317	242

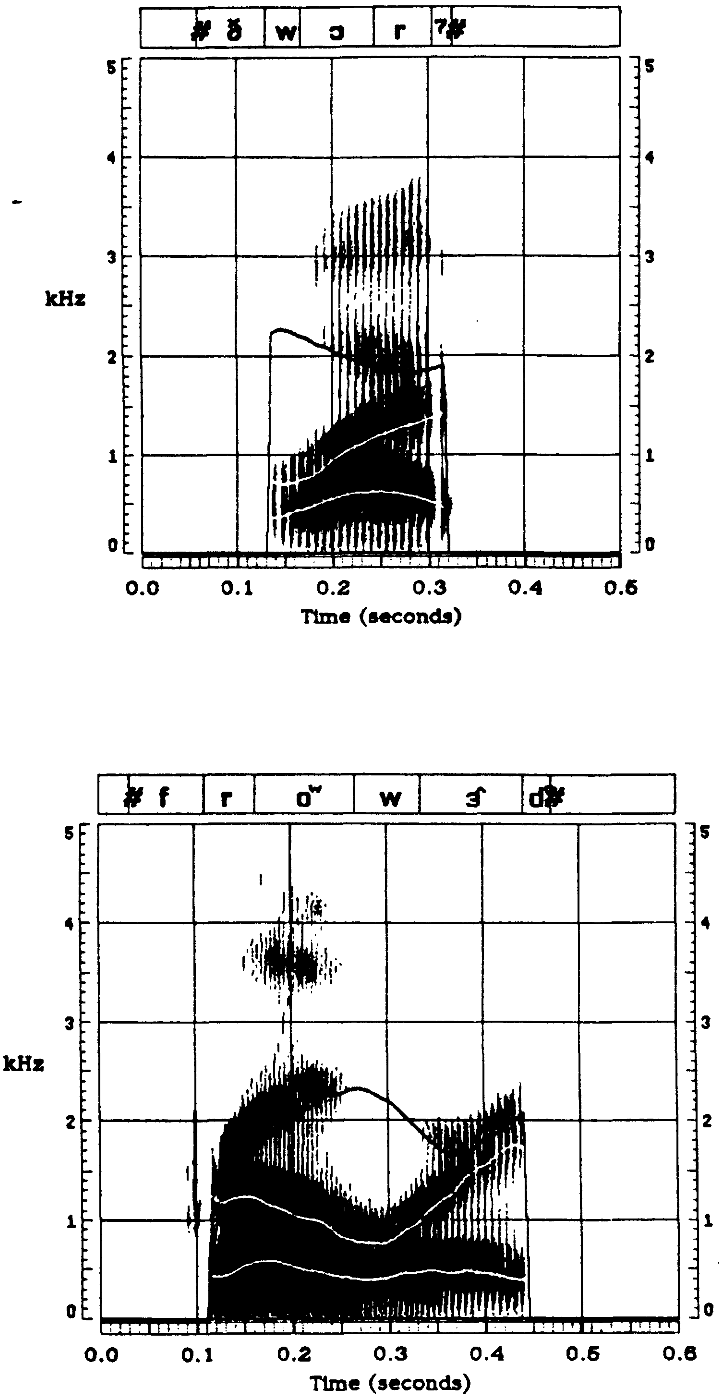


Figure 3.8: An illustration of F3 movement between /w/ and nearby retroflexed sounds in"thwart" and "froward."

/y/

As expected, F1 almost always increases from a /y/ into an adjacent vowel(s), and F2 almost always decreases between /y/ and adjacent vowel(s). Similarly, F3 of a /y/ is normally higher than that of adjacent vowels. There were a few cases where this F3 movement was not observed. In these instances, F3 steadily rose from its value in the /y/ and through the vowel due to the influence of another adjacent consonant, such as the /n/ in "brilliant" ([brɪlyɪnt]) and the /l/ in "uvula"([yuvyulə]).

/l/

As can be inferred from the data, F1 of the vowel is normally higher than F1 of the prevocalic and postvocalic /l/. In the few cases where a postvocalic /l/ had a slightly higher F1 than that of the preceding vowel, the vowel was an /u/. Finally, in the case of an intervocalic /l/, F1 may be lower than F1 of both surrounding vowels, or, due to contextual influences, it may be higher than F1 of one of the surrounding vowels. If /l/ is preceded by a low vowel and followed by a high vowel, such as the second /l/ in "dillydally" ([dɪli^{y}dæli^y]), F1 of /l/ may be higher than F1 of the following high vowel. The converse is true as well. That is, when /l/ is preceded by a high vowel and followed by a low vowel, F1 of the /l/ will sometimes be higher than F1 of the preceding high vowel.

The data also show that, as in the case of /w/, /l/ almost always has a lower F2 frequency than that of the adjacent vowel(s). However, there are a few interesting exceptions which occurred when /l/ was in an intervocalic context. These cases involve the borrowed French words "roulette" and "poilu," spoken by two speakers familiar with the French language. It appears that, in these cases, they produced an /l/ which is different from any /l/ allophones typical of English. Examples of these /l/'s are shown in Figure 3.9.

Finally, the averages and standard deviations of the F3 differences show that F3 almost always increases significantly between /l/ and preceding vowels, and that there is usually little change in F3 between /l/'s and following vowels. These data support previous findings which show that /l/ tends to have an F3 frequency equal to or higher than that of adjacent vowels. However, as can be inferred from the standard deviations, there are several instances where /l/ had a significantly lower F3 frequency than that of the adjacent vowel. This phenomenon, which usually occurs when /l/ is adjacent to a front vowel, can be observed in the words "leapfrog" and "swahili"

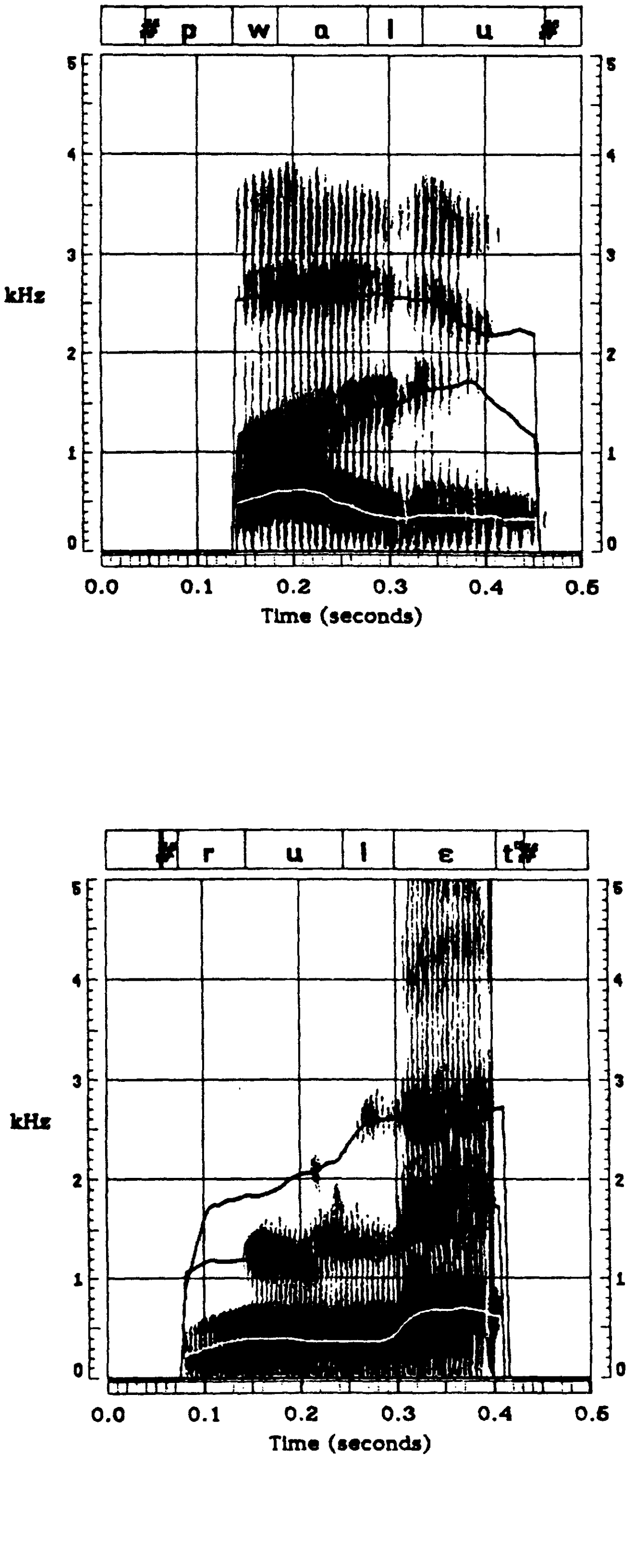


Figure 3.9: Wide band spectrograms of the words "poilu" and "roulette." In both words, /l/ has a higher F2 frequency than an adjacent vowel.

shown in Figure 3.10. Note that, in the latter case, F3 of /l/ is lower than that of both surrounding front vowels.

/r/

When /r/ is in prevocalic position, F1 normally rises from the /r/ into the following vowel. This behavior is often observed between vowels and postvocalic /r/'s as well. However, as can be seen from the data of Table 3.8, there are several instances when the postvocalic /r/ had a higher F1 than that of the preceding vowel. These cases include words like "clear," "queer," "weatherworn" and "yore" where the vowel has a relatively low frequency for F1. Examples of this type of F1 transition are shown in Figure 3.11 for the words "yore" and "clear." Finally, as in the case of /l/, F1 of an intervocalic /r/ may be higher than F1 of one of the adjacent vowels. That is, if /r/ is adjacent to a high vowel and a back vowel, then F1 of the /r/ may lie somewhere between the F1 frequencies of the surrounding vowels.

Between a prevocalic /r/ and a following vowel, F2 may increase or decrease, though the former movement occurs more often. Often, when F2 falls from an /r/ into the following vowel, it is preceded by a coronal consonant such as the /d/ in "withdraw," the /ð/ in "cutthroat" or the /z/ in "Israelite." An example of this type of F2 movement can be observed between the /r/ and /u/ in the word "quadruplet" shown at the top of Figure 3.12. There were a few cases where the F2 differences between the vowel /u/ and the preceding word-initial /r/ in the words "rule" and "roulette" were also negative. However, in all but one case, there was an initial rise in F2 from the /r/ before it fell into its lower value for the /u/. This behavior, which can be seen in the word "rule" also shown in Figure 3.12, was also noted by Lehiste (1962). However, in the word "roulette," shown at the bottom of the figure, this F2 increase is not apparent.

As for intervocalic /r/'s, most have a lower F2 value than adjacent vowels. However, if an intervocalic /r/ is preceded by a back vowel and followed by a front vowel, as in "chlorination" ([klɔrɪne^{y}šən]), then there may be a rise in F2 from the back vowel through the /r/ and into the front vowel. Likewise, if the /r/ is preceded by a front vowel and followed by a back vowel, as in "heroin" ([hɛro^{w}ɪn]), then F2 may fall steadily from the front vowel through the /r/ and into the back vowel.

In the case of postvocalic /r/'s and preceding vowels, F2 may increase or decrease, depending upon whether the vowel is front or back. That is, if the vowel is back, F2

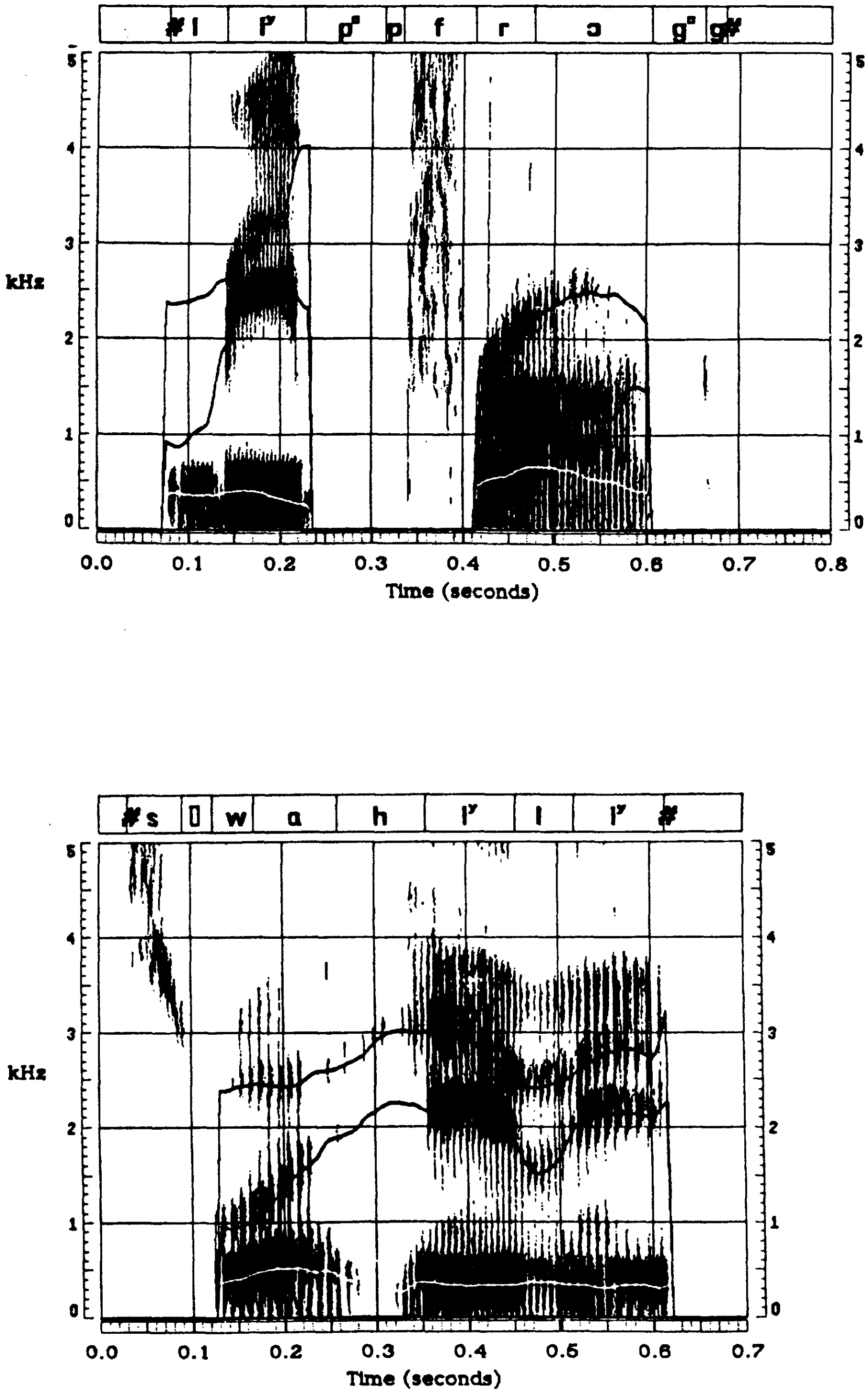


Figure 3.10: Wide band spectrograms of the words "leapfrog" and "swahili." In each case, /l/ has a lower F3 frequency than an adjacent vowel(s).

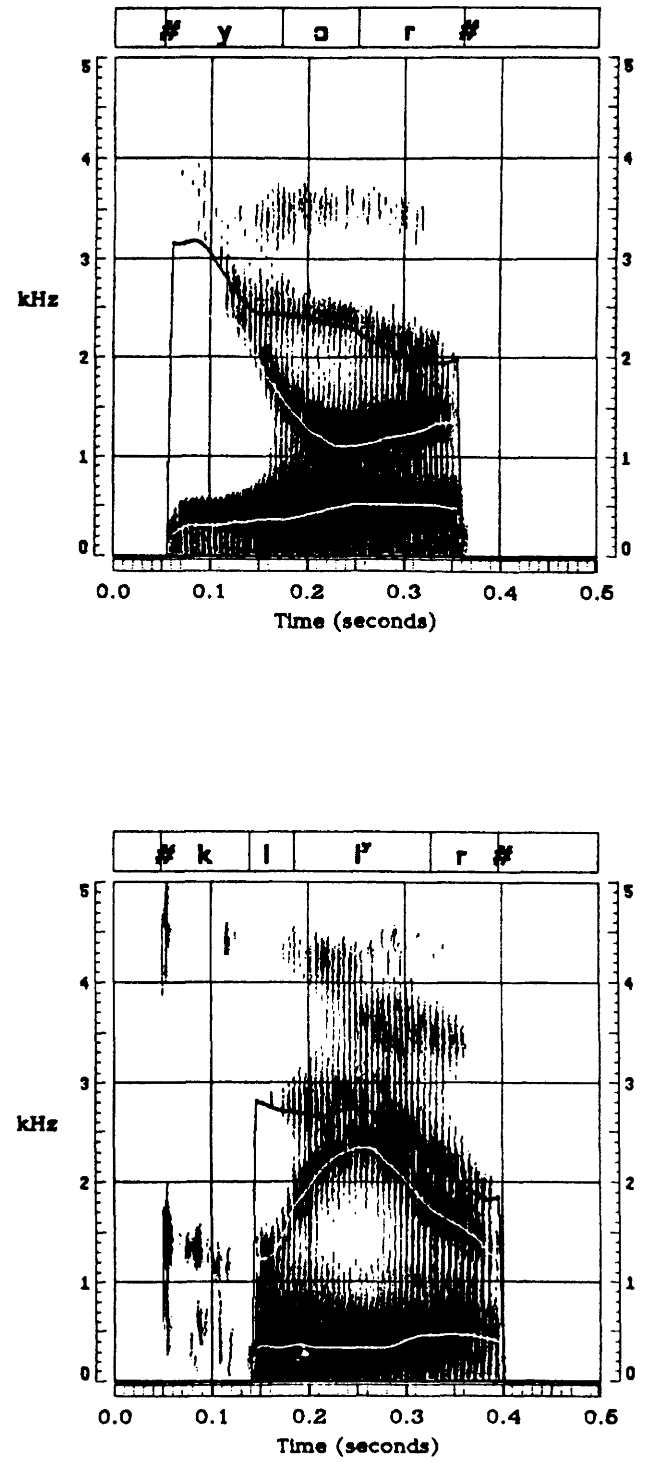


Figure 3.11: Wide band spectrograms of the words "yore" and "clear."

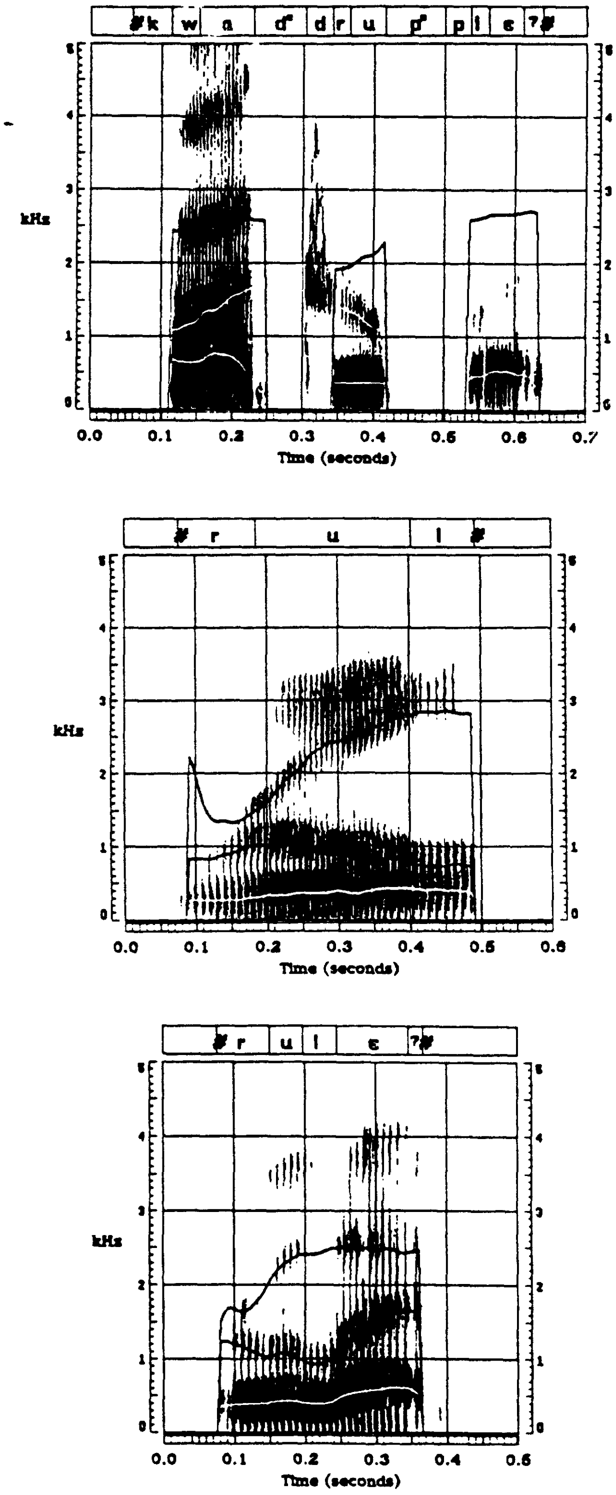


Figure 3.12: Wide band spectrograms of the the words "quadruplet," "rule" and "roulette."

may rise while F3 falls, narrowing the difference between F3 and F2. However, if the vowel is front, both F2 and F3 will fall into the appropriate values for an /r/. This behavior can also be observed in the words "yore" and "clear" shown in Figure 3.11.

As expected, F3 almost always increases between a prevocalic /r/ and the following vowel. However, there was a notable exception. This case involved the the word "rauwolfia" which, instead of being pronounced as [rɔwo^{w}lfi^{y}ɑ], was pronounced as [rɔro^{w}lfi^{y}ɑ]. That is, the speaker replaced the intervocalic /w/ with an intervocalic /r/. Due to the influence of the intervocalic /r/, F3 falls by 220 Hz between the prevocalic /r/ and the /ɔ/. This behavior can be observed in Figure 3.13, where F3 steadily decreases from its first visible value within the word-initial /r/ to its lowest value within the intervocalic /r/. This behavior, which is observable from the formant tracks which are extracted within the portion where F3 is not visible on the spectrogram, was verified from wide-band and narrow-band short-time spectra. This is the type of F3 movement we would expect to see between prevocalic /w/'s and following retroflexed sounds. However, when this utterance is played, a clear word-initial /r/ is heard.

In the case of intervocalic /r/, F3 is almost always equal to or lower than that of adjacent vowels. There was an exception which occurred in the word "guarani," shown in Figure 3.14. In this case, the vowel /ɑ/ preceding the /r/ is retroflexed so that the lowest point of F3 is within the vowel region.

Finally, the data of Table 3.8 show that postvocalic /r/'s generally have a lower F3 value than the preceding vowel. However, as can be inferred by the large standard deviation, there are some instances where a postvocalic /r/ has a higher F3 value than that of the preceding vowel. This behavior was observed only in words where the /r/ is not in word-final position, but is followed by another consonant, such as the /r/'s in "cartwheel," "harlequin" and "Norwegian." Furthermore, as was seen in the example of the word "guarani," there is significant feature assimilation between the vowel and the following /r/ such that the vowel is retroflexed throughout. In these cases, the lowest point of F3 within the syllabic region can occur near the beginning of the vowel. This phenomenon is discussed further in Section 3.3.

3.2.3 Relative Low-Frequency Energy Measures

As we stated earlier, the production of the semivowels is in many ways similar to the production of vowels. The vocal folds vibrate during the articulation of the

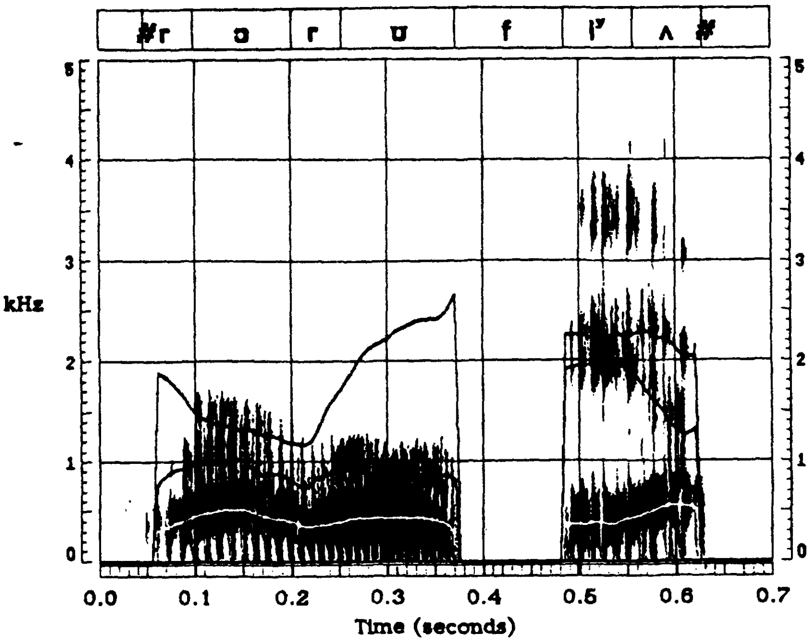


Figure 3.13: Wide band spectrogram with formant tracks overlaid of the word "rauwolfia" where the intervocalic /w/ was replaced by an intervocalic /r/. Note the downward movement from the word-initial /r/ and the intervocalic /r/.

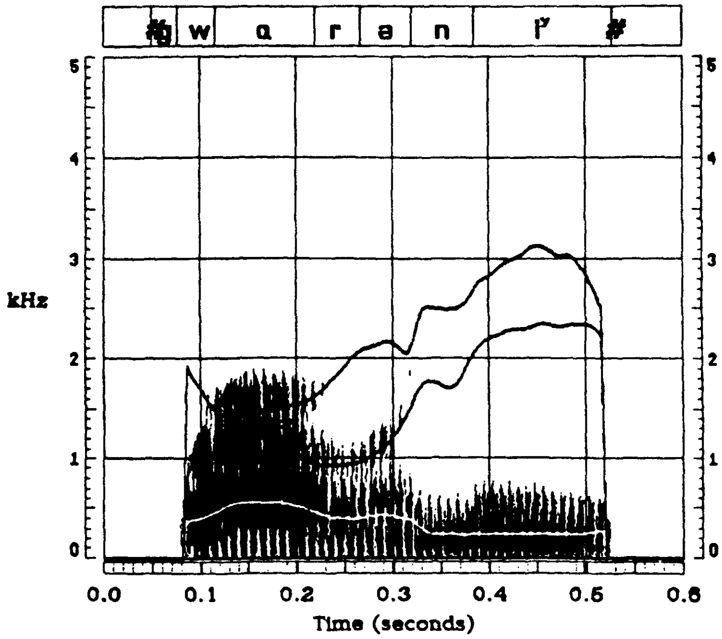


Figure 3.14: Wide band spectrogram of the word "guarani." The lowest point of F3 occurs during the retroflexed vowel /ɑ/.

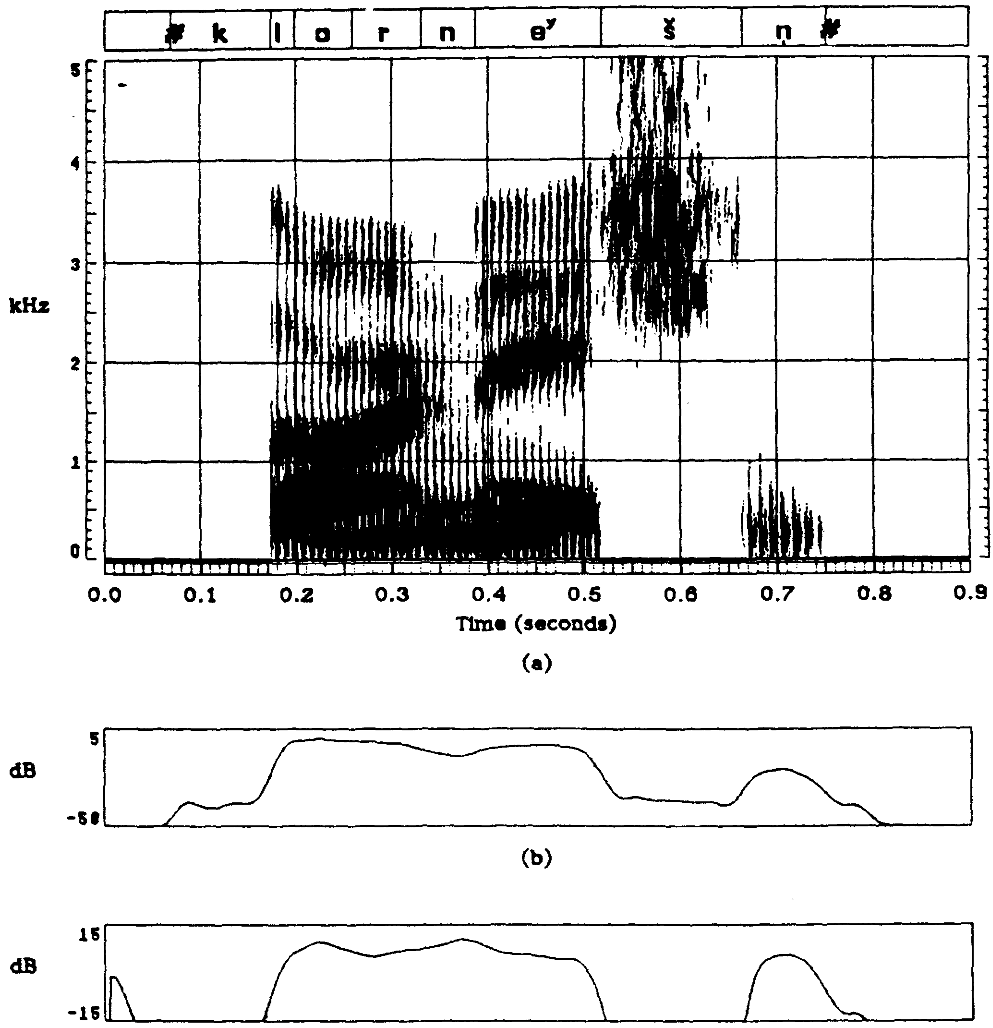
semivowels and, unlike many consonants, no frication noise is produced. The only other consonants which share these properties are the nasals. Hence, the semivowels, vowels and nasals are considered to be voiced sonorant sounds.

In this part of the acoustic study, we attempted to determine robust acoustic correlates of these *voiced* and *sonorant* features. The acoustic correlate normally used for the feature *voiced* is low frequency periodicity. However, the available pitch tracker (Gold and Rabiner, 1969) does not always accurately estimate the beginning of voiced and sonorant segments. Therefore, we used a low-frequency energy measure instead. This energy measure is based on the bandlimited energy computed from 200 Hz to 700 Hz. More specifically, the value of the parameter in each frame is the difference (in dB) between the maximum energy within the utterance and the energy in each frame. An example of this parameter is shown in part b of Figure 3.15 for the word "chlorination." As can be seen, the energy difference is small in the vowel, semivowel and nasal regions, and large and negative in value in the stop and fricative regions.

The parameter used to capture the feature ***sonorant*** is the difference (in dB) between the high-frequency energy computed from 3700 Hz to 7000 Hz and the low-frequency energy computed from 100 Hz to 300 Hz. Thus, for vowels, nasals and semivowels, which have considerable low-frequency energy and some high-frequency energy, this difference should be small. However, for nonsonorant consonants, like fricatives which have mainly high-frequency energy, this difference should be high. This behavior can be seen in part c of Figure 3.15.

The results obtained with these parameters are shown in Figure 3.16. Separate scatter plots are shown for the vowels, the nasals and semivowels, and the remaining consonants. Statistical data concerning the averages and standard deviations are also given.

As can be seen, there is almost complete overlap between the vowels (about 2400 tokens), and the semivowels and nasals (about 2200 tokens). However, there is very little overlap between these voiced sonorant sounds and the remaining consonants (about 2400 tokens). Only about 16% of the remaining consonants overlap with the voiced sonorant sounds. Of these overlapping consonants, 79% are voiced consonants, including flaps, glottal stops, fricatives, stops and affricates. Excluding the glottal stops (which make up one fourth of these voiced consonants), 71% of the voiced consonants are in intervocalic position or, more generally, in intersonorant (between two sonorants) position. Spectrograms of words containing two of these consonants, the



(a)

(b)

(c)

Figure 3.15: An illustration of parameters used to capture the features *voiced* and *sonorant.* (a) Wide band spectrogram of the word "chlorination." (b) Energy difference (100 Hz to 700 Hz) between maximum value in utterance and value in each frame. (c) Difference between low-frequency energy (100 Hz to 300 Hz) and high-frequency energy (3700 Hz and 7000 Hz).

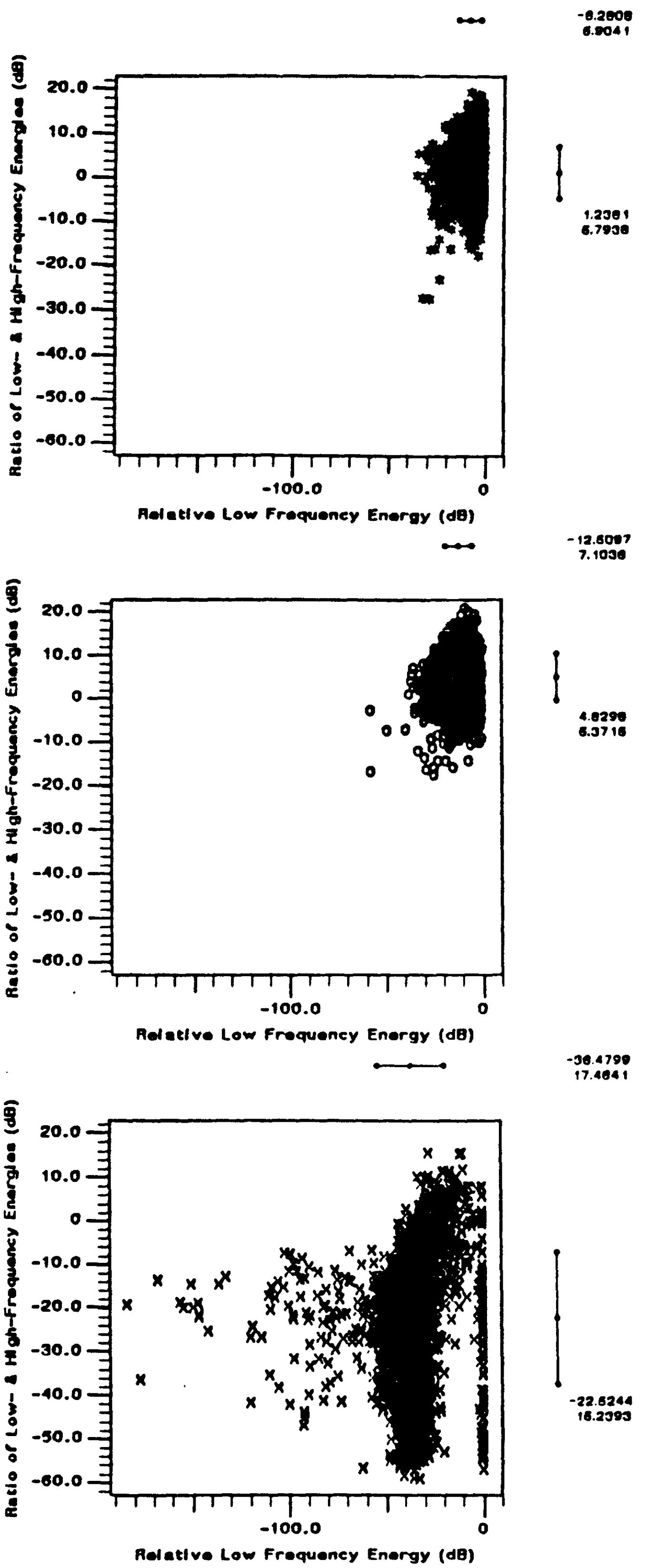


Figure 3.16: Results obtained with the voiced and sonorant parameters. vowels: * semivowels and nasals: o, other consonants: ×.

intervocalic /g/ in "wagonette" and the intervocalic /v/ in "wolverine," are shown in Figure 3.17. As can be seen, the intervocalic /g/ has no burst or voice onset time. Instead, the /g/ segment appears to be sonorant throughout. Likewise, the intervocalic /v/, which has no frication noise, also appears to be sonorant throughout. Thus, the feature *sonorant*, which is generally absent from voiced stops, fricatives and affricates, is sometimes shared by these sounds when they are surrounded by sonorant segments.

Many of the remaining nonintervocalic and voiced consonants that overlap with the vowels, nasals and semivowels, are unreleased stops, which occur in word-final position. Overlapping prevocalic stops usually occur before back or retroflexed sounds such that they have low-frequency bursts. This latter phenomenon can be observed for the /g/ burst in the word "granular," shown at the top of Figure 3.18. The nonintervocalic and voiced fricatives which overlap with the sonorants are all /v/'s that occur mainly in word-final position. An example of such a /v/ occurs in the word "exclusive," also shown in Figure 3.18. Note that the word-final /v/ is very weak and has no frication noise.

Finally, those unvoiced stops which overlap with the semivowels, vowels and nasals are either unreleased and in word-final position, or they occur in prevocalic position before back sounds such that they have low-frequency bursts. Such a stop is the /k/ in the word "queen" shown at the bottom of Figure 2.3 of Chapter 2.

In summary, the results of this section show that, in addition to the semivowels and nasals, other voiced consonants may appear as sonorant in certain environments. However, with these parameters, a few nonsonorant consonants are confused with the sonorant sounds.

3.2.4 Mid-Frequency Energy Change

Vowels, because they are less constricted, usually have considerably more energy in the low- to mid-frequency range than the semivowels and other consonants. That is, the semivowels, like other consonants, usually occur at syllable boundaries. A syllable boundary can be defined acoustically as a significant dip within some bandlimited energy contour. To access this difference in energy between semivowels and vowels, and, more generally, between consonants and vowels, we used two bandlimited energies in the frequency ranges 640 Hz to 2800 Hz and 2000 Hz to 3000 Hz.

We chose the frequency range 640 Hz to 2800 Hz because, relative to the vowels, the semivowels tend to have less energy in this region. This can be seen in Figure 3.19

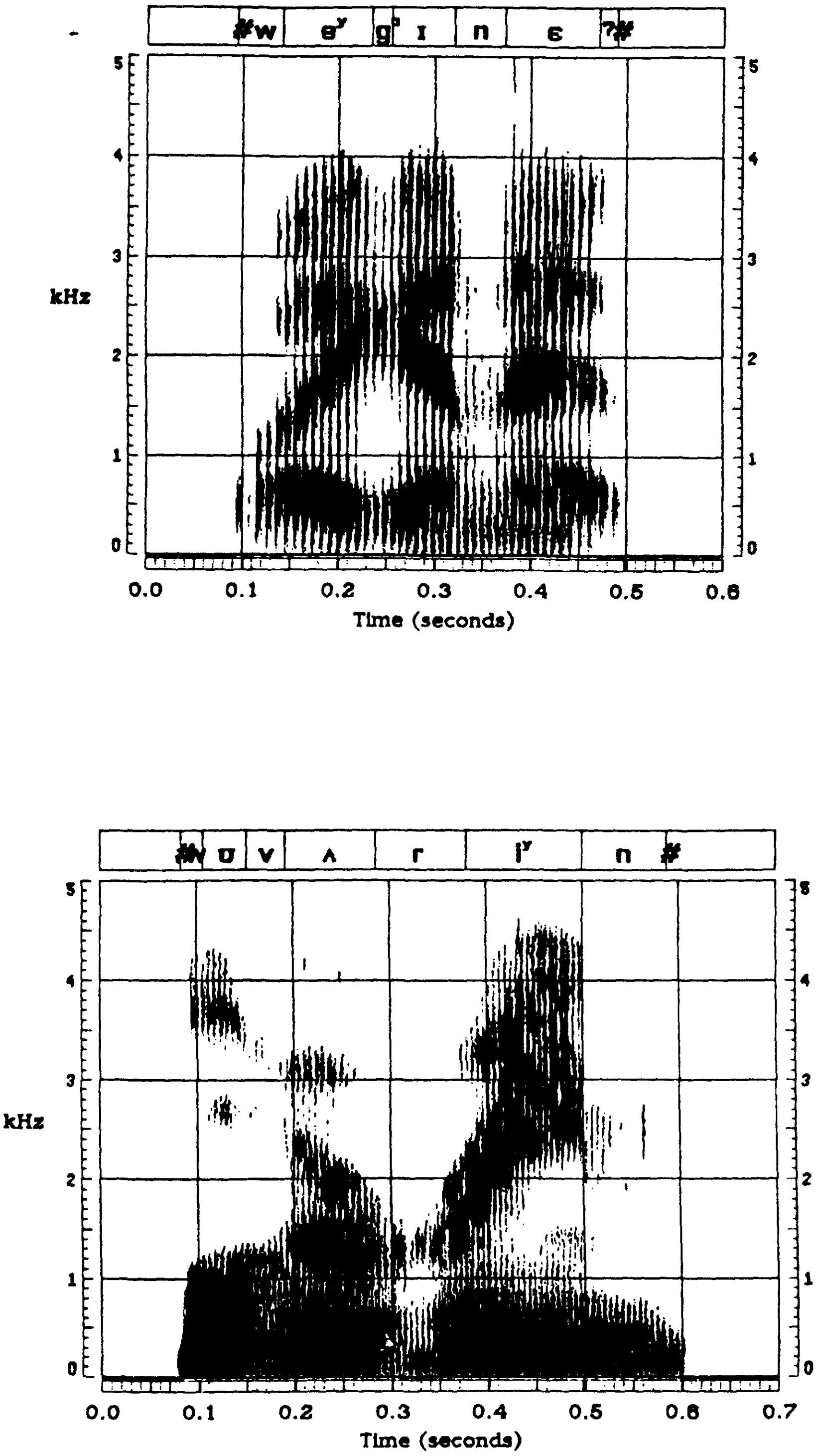


Figure 3.17: Wide band spectrogram of the word "wagonette" which contains a sonorant-like /g/ and "wolverine" which contains a sonorant-like /v/.

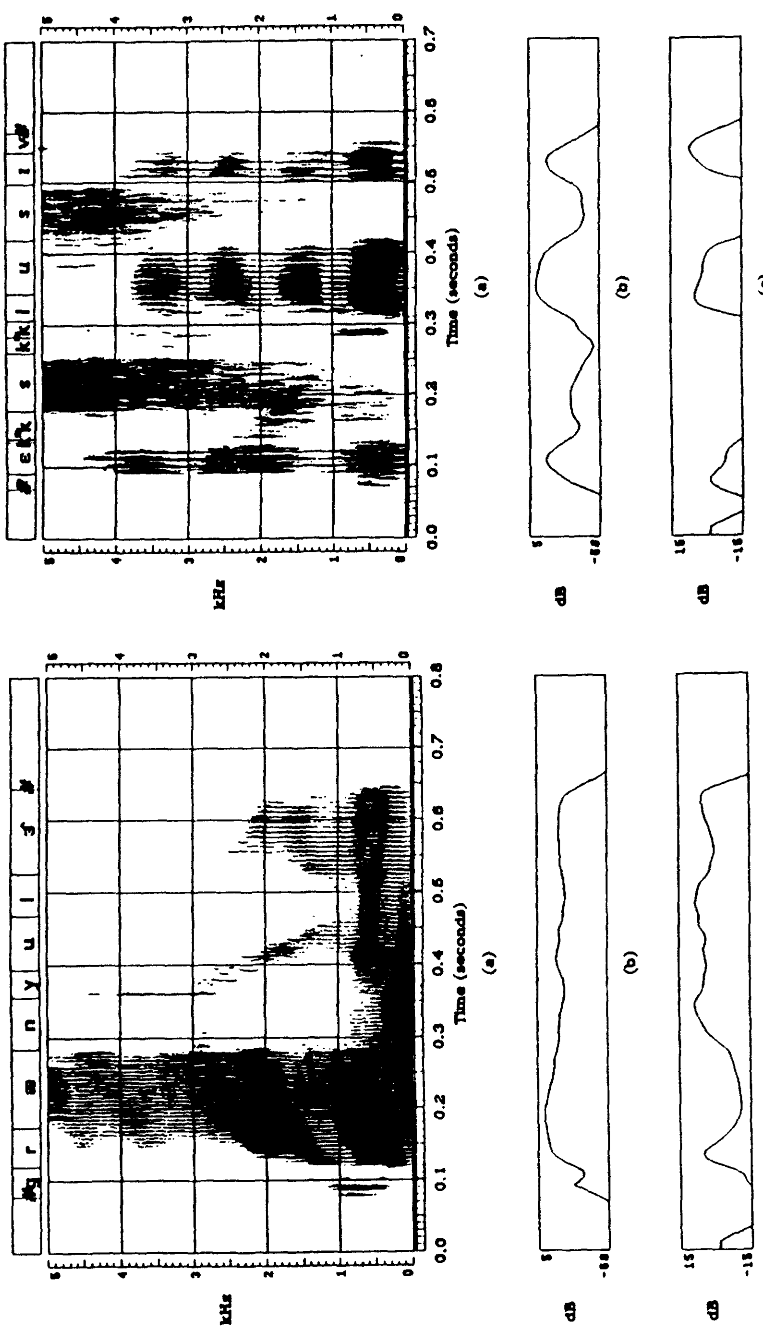


Figure 3.18: Voiced and Sonorant parameters of the words "granular" (left) and "exclusive" (right). (a) Wide band spectrograms. (b) Difference (in dB) in energy (100 Hz to 700 Hz) between maximum value in utterance and value in each frame. (c) Difference (in dB) between low-frequency energy (100 Hz to 300 Hz) and high-frequency energy (3700 Hz and 7000 Hz).

for the /w/ in "periwig," the /r/ in "diuretic," and the /l/ and /y/ in "humiliate." The lower bound of 640 Hz will usually exclude F1 which, for most sounds, is stronger than F2 and F3. In addition, as can be seen in "humiliate," the amplitude of F1 for the semivowels can be comparable to that of the adjacent vowels. The upper limit of 2800 Hz includes F2 and F3 which, relative to the vowels, tend to be weaker for the semivowels. The low amplitude of F3 for /w/ is probably due to its very low F1 and F2 frequencies. This acoustic property supports perceptual results obtained by O'Connor et al. (1957). They found that an acceptable /w/ could be synthesized using only the first two formants. In fact, including F3 in the synthesis made little or no perceptible difference. A weak third formant is also characteristic of the semivowel /l/. In this case, the low amplitude formant is due to a close lying antiresonance caused by the shunting effect of the mouth cavity behind the tongue blade (Fant, 1960).

The frequency range 2000 Hz to 3000 Hz was chosen to aid in the detection of /r/'s. From a preliminary study, we found that many intervocalic /r/'s had energy, in the frequency range 640 Hz to 2800 Hz, comparable to that of adjacent vowels. Such an /r/ is in the word "periwig" shown in Figure 3.19. As can be seen, F1, F2 and F3 of the /r/ are all strong. However, since F3 is normally between 2000 Hz and 3000 Hz for vowels, but falls near or below 2000 Hz for /r/, /r/ will usually be considerably weaker in the 2000 Hz to 3000 Hz range than an adjacent vowel(s).

We discuss separately below the effectiveness of these bandlimited energies in identifying the presence of semivowels and other consonants when they occur in intervocalic, prevocalic and postvocalic contexts.

Intervocalic Consonants

For each of the energy parameters, the difference (in dB) between the minimum energy within the semivowels and other consonants, and the maximum energy within the adjacent vowels was measured. The smaller of these two differences determines the depth of the energy dip. An example of this measurement is given in Figure 3.20 for the word "bewail." As can be seen from the bandlimited energy waveform shown in part b of Figure 3.20, an energy dip of 28 dB occurs within the intervocalic /w/ at about 190 msec.

To determine if similar energy dips occurred within vowels, we used a similar measurement procedure illustrated in Figure 3.21 for the word "yon." Within the hand-transcribed vowel region, we made several measurements. First, we determined

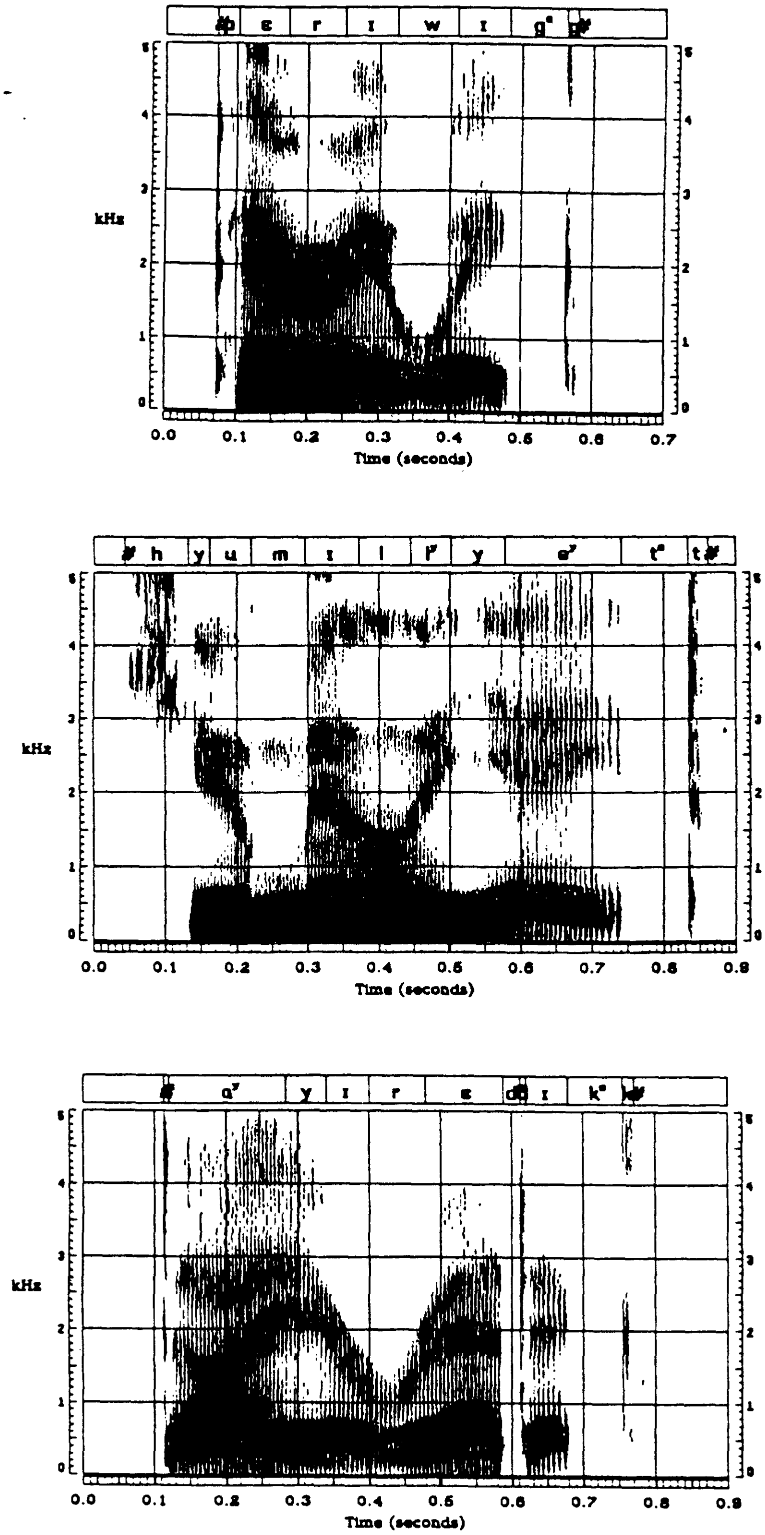


Figure 3.19: Wide band spectrogram of the words "periwig," "humiliate" and "diuretic."

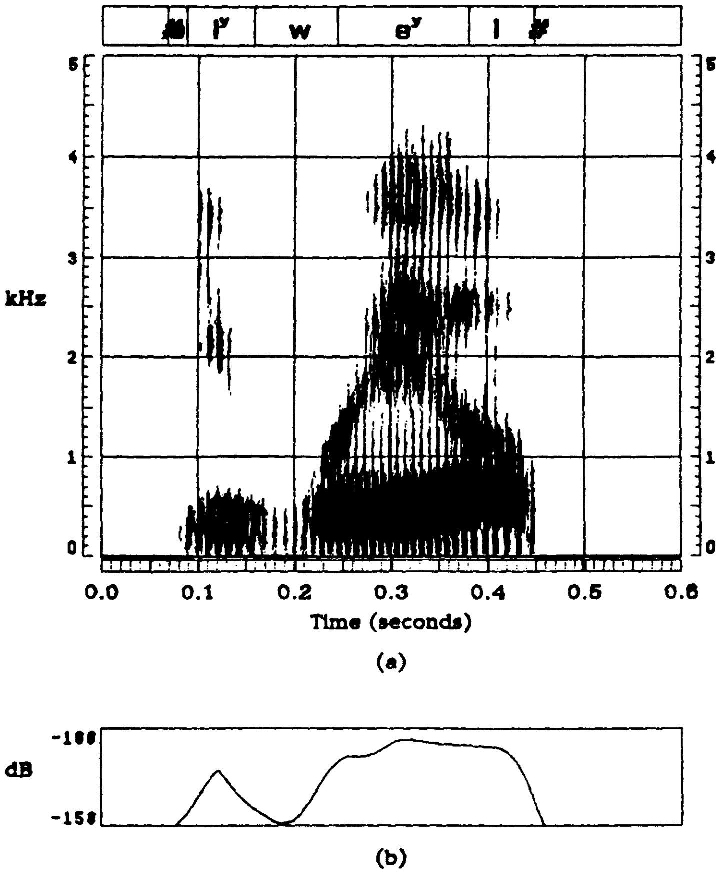
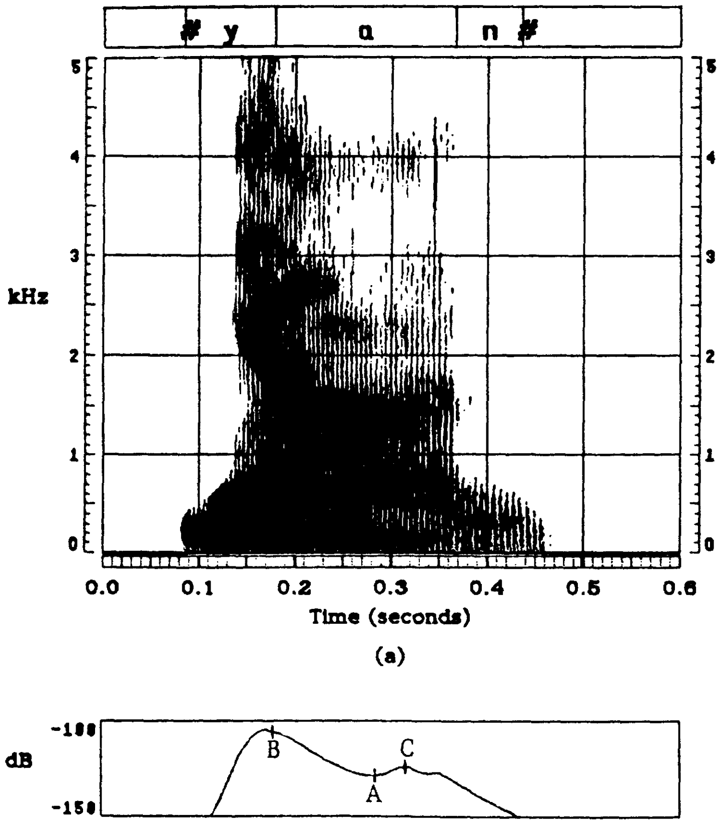


Figure 3.20: Measurement procedure for energy dips within intervocalic consonants. (a) Wide band spectrogram of the word "bewail." (b) Energy 640 Hz to 2800 Hz.



(a)

(b)

Figure 3.21: Measurement procedure for intravowel energy dips. (a) Wide band spectrogram of the word "yon." (b) Energy 640 Hz to 2800 Hz.

the minimum energy and the time at which it occurs. This instant of time is marked as point A in part b of Figure 3.21. Second, we determined the maximum energy between the beginning of the vowel region and point A. The frame in which this maximum energy occurs is marked as point B. Finally, we determine the maximum energy occurring between point A and the end of the vowel region. The frame at which this maximum energy occurs is marked as point C. The smaller of the differences in energy at times B and A and at times C and A determines the depth of the intravowel energy dip. In this example, the depth of this dip is 4 dB.

The results of the above measurements are shown in Figure 3.22. In part a, which contains measurements made on about 2400 vowels, we see that usually there is no intravowel energy dip. In most instances where there is a significant intravowel energy dip larger than 2 dB, the vowel is an /ɝ/ or a diphthong. For example, consider the /ɝ/ in the word "plurality" and the $/i^y/$ in the word "queer," shown in Figure 3.23. In both instances, portions of the transcribed vowels appear to be nonsyllabic. Although no clear /r/ and /y/ were heard, their exclusion from the transcription is questionable.

Most of the nonsonorant and nasal consonants shown in parts b and c of Figure 3.22 have significant energy dips in one or both bandlimited energies. Those consonants which have as much or more energy than the adjacent vowels are the strong fricatives /š/ and /ž/, which have considerable energy in the range 2000 Hz to 3000 Hz. Recall that the speech signals were preemphasized.

Finally, the results for the semivowels, which are shown in part d of the figure, show that they usually have significantly less energy than the surrounding vowels. However, 10% of the semivowels did not have a significant ($\geq$ 2 dB) energy dip in either of the bandlimited energies. More specifically, 33% of the /y/'s, 14% of the /r/'s and 5% of the /l/'s did not contain significant energy dips.

On close examination of the semivowels which do not appear to be nonsyllabic, certain patterns emerged. In nearly all of the words containing either an intervocalic /l/ or /r/ with no energy dip, the /l/ or /r/ followed a stressed vowel and preceded an unstressed vowel, such as the /l/'s in "swollen," "plurality" and "astrology," and the /r/'s in "heroin," "marijuana" and "guarantee." There was, however, an exception which involved the /l/ in "musculature," where the /l/ followed an almost devoiced /ə/. Examples of one of the /l/'s and one of the /r/'s are shown in Figure 3.24.

The case of the intervocalic /y/'s that do not contain significant energy dips is more complicated. In 12 out of 14 words containing a /y/ with no significant energy dip, the /y/ segment is a result of the offglide of a diphthong, such as the /e^y/ in "humiliate" and the /ɔ^y/ in "flamboyant." The two cases where this was not the case involved the words "volume" (pronounced as [vayum]) and "cellular" (pronounced as [sɛyulɚ]).

As in the case of /l/ and /r/, 64% of the /y/'s with no significant energy dip preceded vowels with less stress then the vowels they followed, such as the /y/'s in the words "brian" and "diuretic." The exceptions to this pattern involved the words "radiology," "humiliate," "unreality" and "riyal."

From a reexamination of these words, we found that a clear /y/ was heard in most of them when we played either the entire utterance or some portion thereof. A comparison of the words "humiliate," which contains a clearly heard /y/, and "Ghanaian," which contains a questionable /y/, are shown in Figure 3.25.

It is not clear what we should conclude about this lack of significant energy dips within 10% of the intervocalic semivowels. It may be that some syllable boundaries

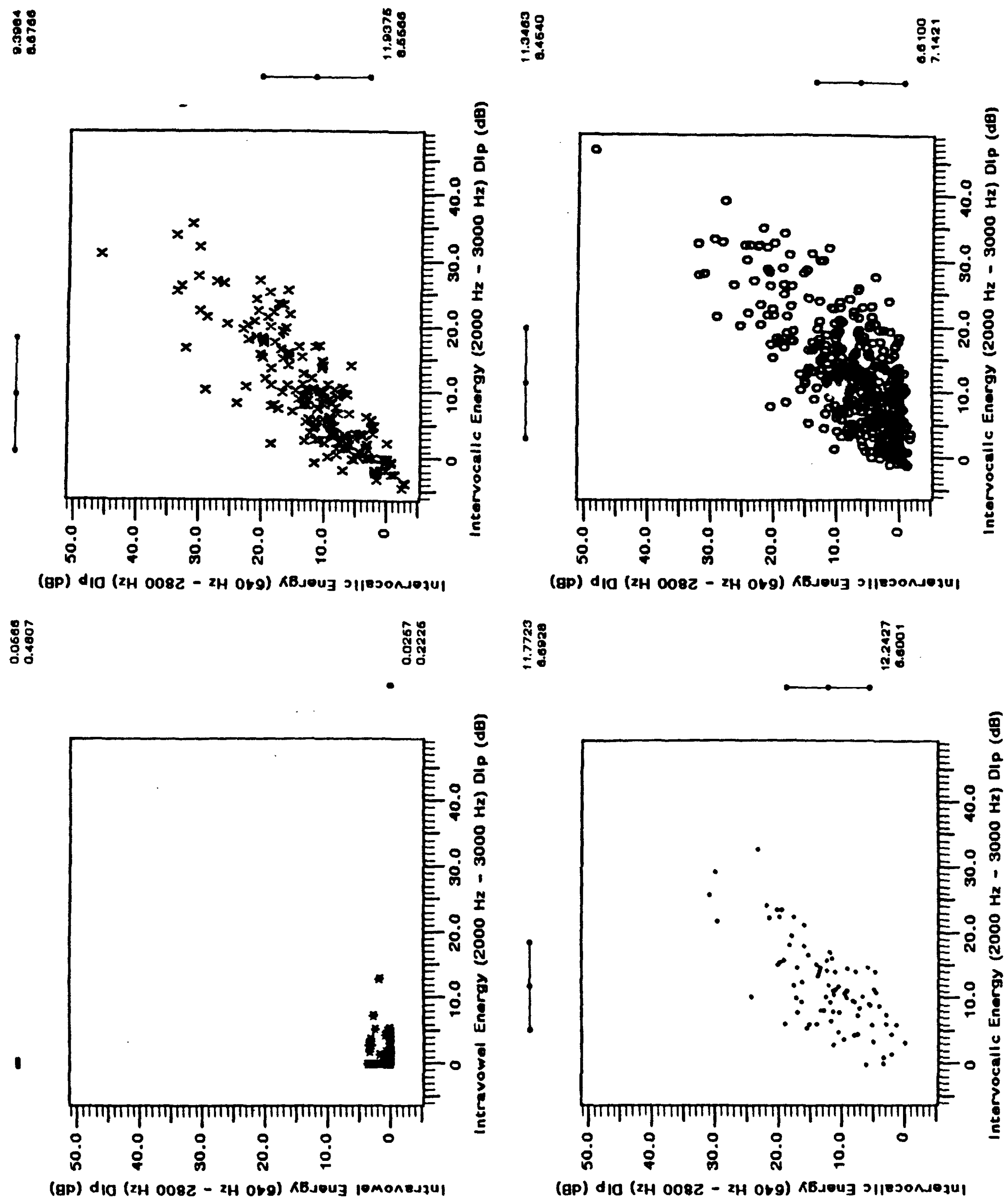


Figure 3.22: Comparisons between intravowel energy dips and average energy differences between intervocalic consonants and adjacent vowels. vowels: *, nonsonorant consonants: x, nasals: ., semivowels: o.

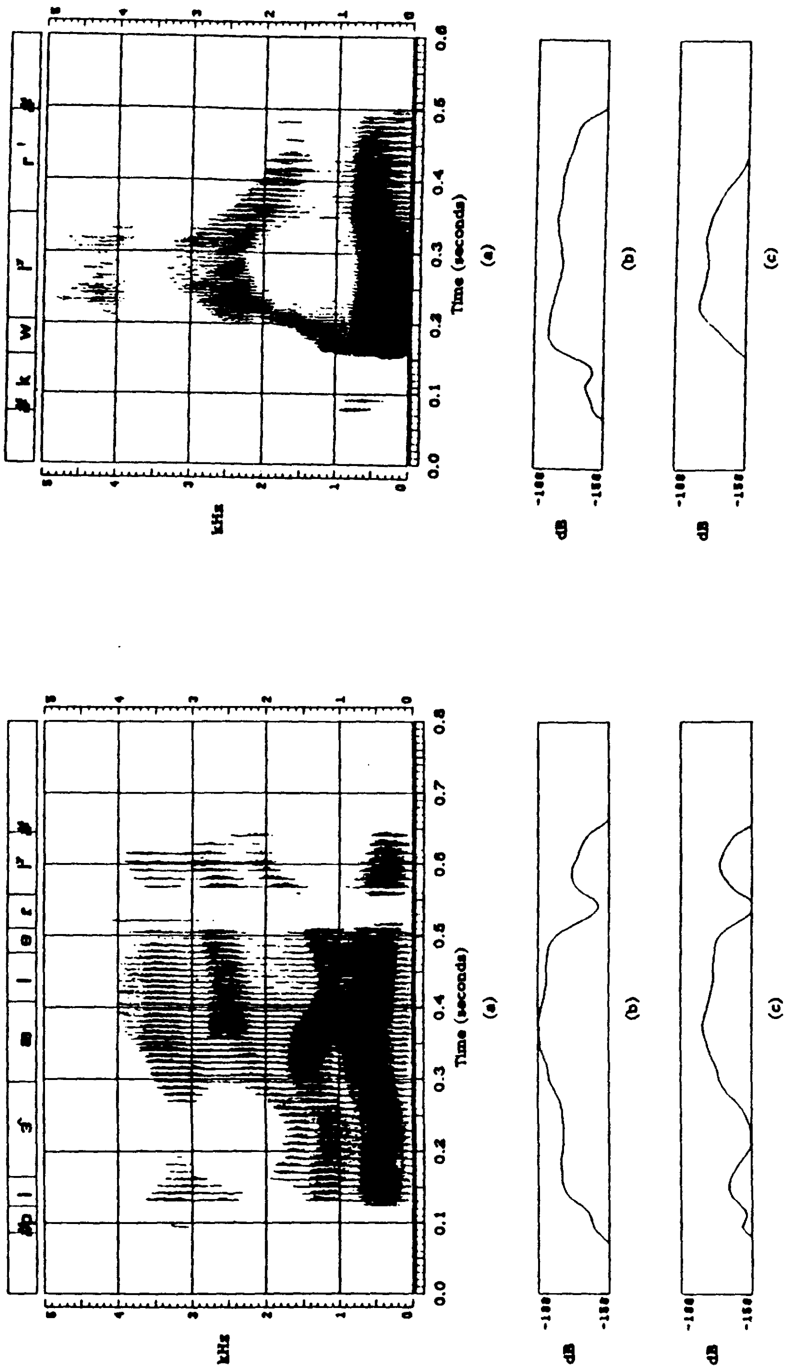


Figure 3.23: Significant intravowel energy dips. (a) Wide band spectrograms of "plurality" and "queer." (b) Energy 640 Hz to 2800 Hz. (c) Energy 2000 Hz to 3000 Hz.

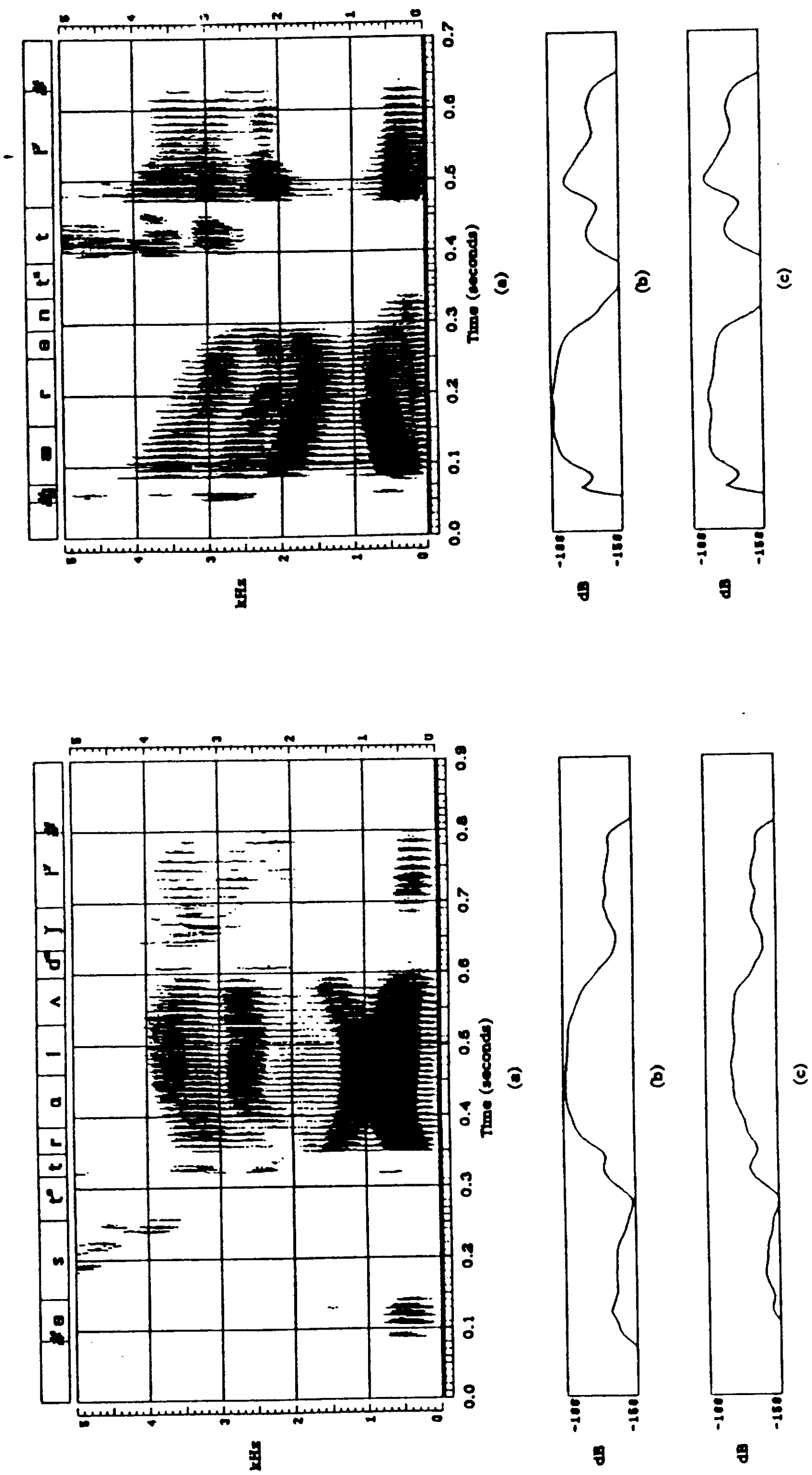


Figure 3.24: Intervocalic semivowels with no significant energy dips. (a) Wide band spectrograms of "astrology" and "guarantee." (b) Energy 640 Hz to 2800 Hz. (c) Energy 2000 Hz to 3000 Hz.

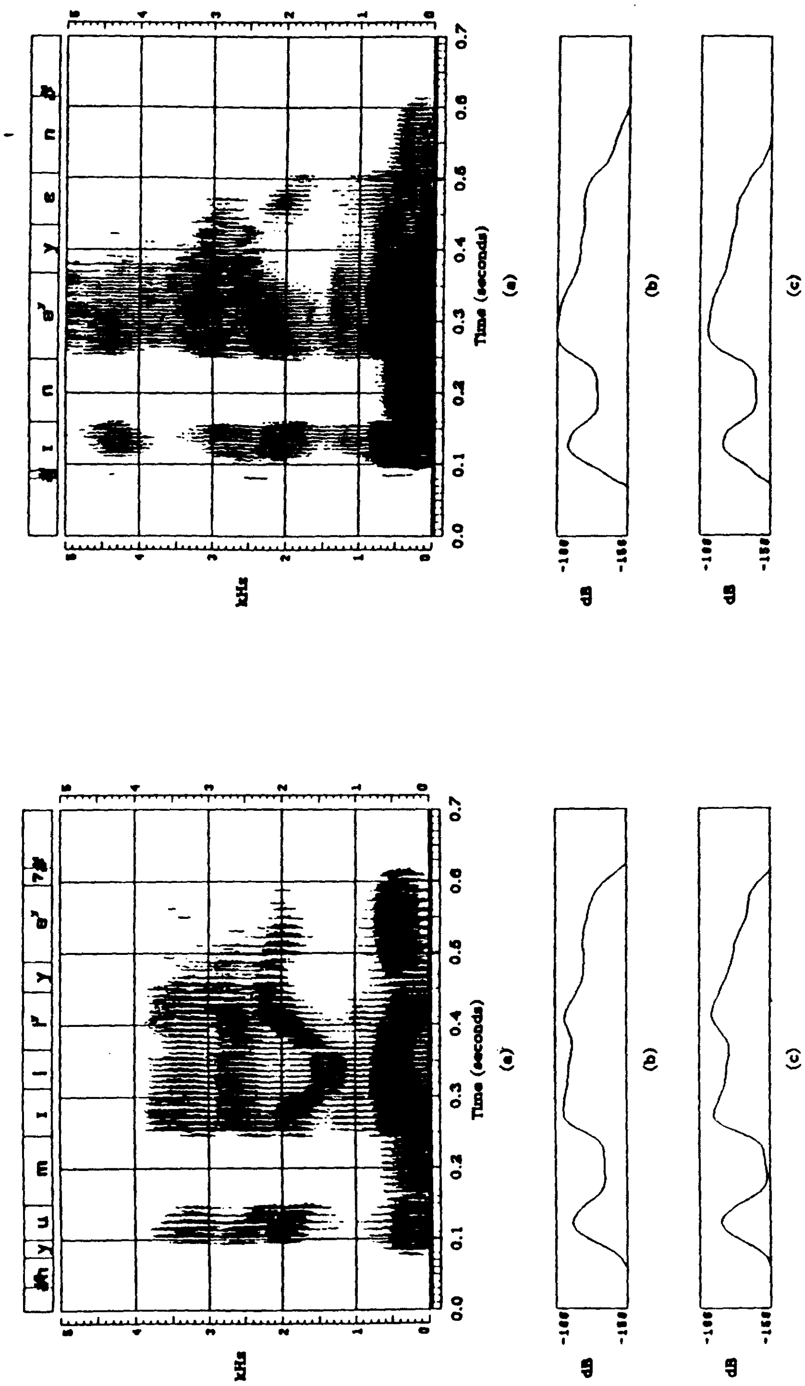


Figure 3.25: Intervocalic /y/'s with no significant energy dips. (a) Wide band spectrograms of "humiliate" and "Ghanaian." (b) Energy 640 Hz to 2800 Hz. (c) Energy 2000 Hz to 3000 Hz.

are perceived only after we extract words from our lexicon. It may be that some additional acoustic property(s), such as formant movement, helps us to perceive syllable boundaries. Or, it may be that some other bandlimited energy would result in their detection. For example, since /y/ normally has F2 and F3 frequencies above 2000 Hz, a bandlimited energy computed from 1000 Hz to 2000 Hz may contain dips within more of the /y/ segments. Clearly this phenomenon needs to be studied further.

Prevocalic Consonants

To ascertain the effectiveness of the bandlimited energies in identifying the prevocalic semivowels and other consonants, we compared the minimum energy within the consonants (the beginning of the consonant region was taken to be the smaller of either 10 msec or 20% into the hand-transcribed consonant region) with the maximum energy within the following vowel. For comparison, we also measured the depth of similar energy changes occurring naturally within word-initial vowels. An example of the latter measurement procedure is given in Figure 3.26 for the word "always." First, we compute the maximum energy within the vowel and the time at which it occurs. This frame is labeled point A in part b. Second, between the beginning of the vowel (starting at the smaller of 10 msec or 20% into the hand-transcribed vowel region) and point A, we compute the minimum energy and the time at which it occurs. This frame is labeled point B. The difference (in dB) between the maximum energy and minimum energy at these times is the depth of the intravowel energy dip. For the /ɔ/ in the example, the intravowel energy dip is 11 dB.

The results for the vowels and consonants are compared in Figure 3.27. As can be seen in part a, the average energy increase within vowel regions is about 12 dB. However, the energy can increase by as much as 30 dB. The average increase in energy between nonsonorant consonants and vowels and between nasals and vowels is between 28 dB and 33 dB. Between the semivowels and following vowels, the average energy increase is about 21 dB, and 40% of the semivowel-vowel transitions involve an energy increase of more than 30 dB. If we look only at the energy change between word-initial semivowels and following vowels, the average energy increase is about 30 dB, and 62% of the semivowel-vowel transitions have an energy increase of more than 30 dB.

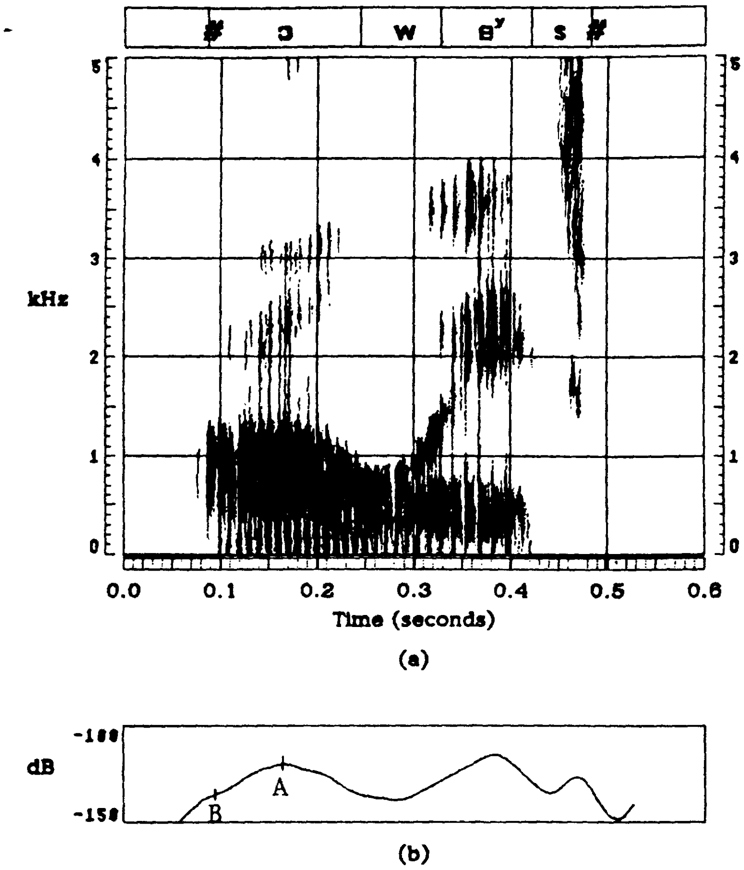


Figure 3.26: Measurement procedure for natural energy increase in word-initial vowels. (a) Wide band spectrogram of the word "always." (b) Energy 640 Hz to 2800 Hz.

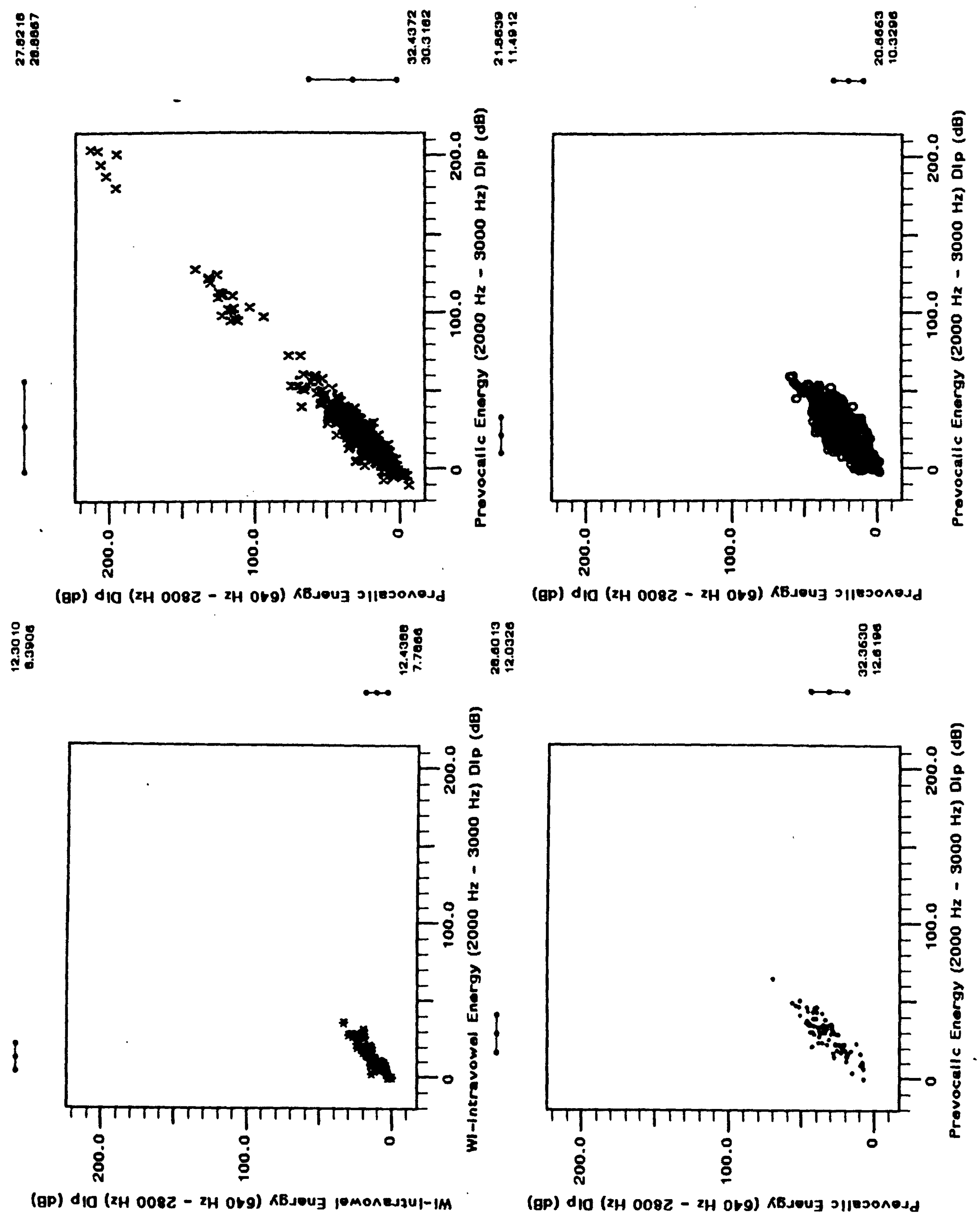


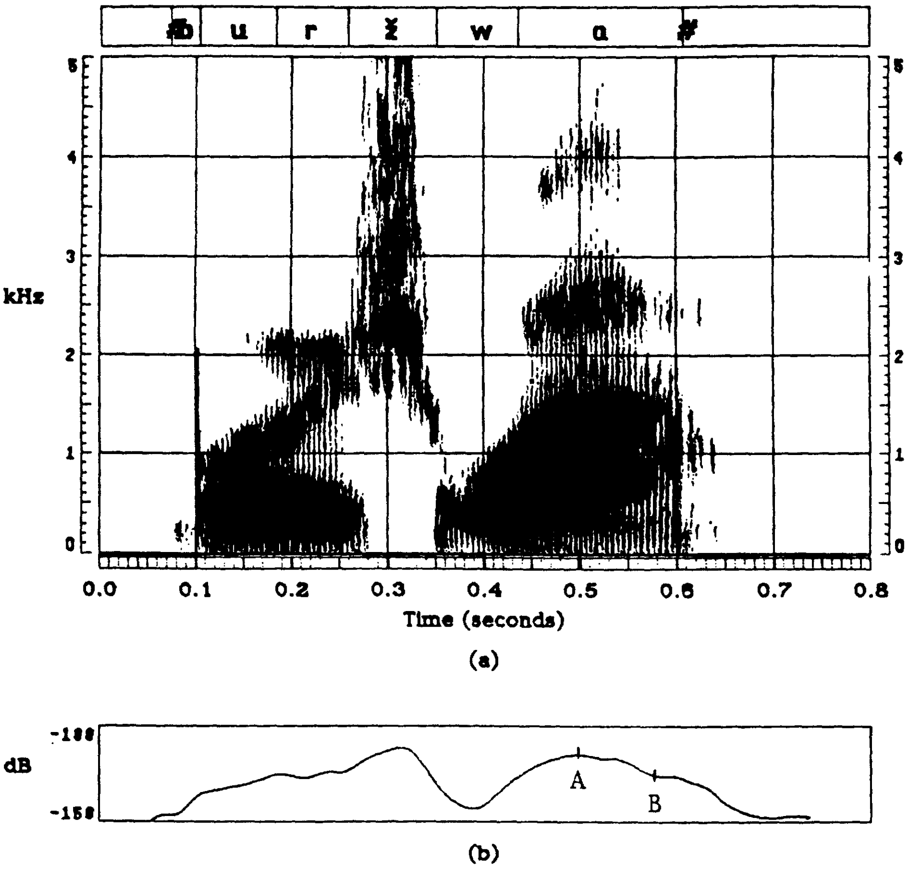
Figure 3.27: Comparisons of natural energy rise within vowels and average energy difference between prevocalic consonants and following vowels. vowels: *, nonsonorant consonants: x, nasals: ., semivowels: o.

Postvocalic Consonants

To determine the depth of energy dips occurring between postvocalic consonants and preceding vowels, we computed the difference (in dB) between the maximum energy within the vowel regions and the minimum energy within the postvocalic consonant (where the end of the consonant region is considered to be the larger of 10 msec before the end of the hand-transcribed region or 80% of the hand-transcribed region). For comparison, we measured the natural decrease in energy within word-final vowels. This measurement procedure for the vowels is illustrated in Figure 3.28 for the word "bourgeois." First, we determined the maximum energy and the time at which it occurs. This frame is labeled point A in part b. Second, we compute the minimum energy occurring between this time and the end of the vowel region (where the end of the vowel region is the larger of 10 msec before the end of the hand-transcribed region or 80% of the hand-transcribed energy). This frame is labeled point B. The intravowel energy dip is taken to be the difference between the maximum and minimum energy. In this example, the intravowel energy dip is 14 dB.

The distributions of energy dips occurring within word-final vowels and between vowels and postvocalic consonants are shown in Figure 3.29. As can be seen in part a, the vowels have an average natural energy taper between 12 dB and 14 dB. Most of the vowels with an energy dip of more than 20 dB are diphthongs. That is, a large energy change is usually due to a /y/ or /w/ offglide. An example of this significant decrease in energy is shown in Figure 3.30 for the word "view," which has a 50 dB energy dip in the frequency range 2000 Hz to 3000 Hz. If we exclude diphthongs and syllabic nasals from the word-final vowels, the average energy dip drops to 11 dB with a maximum energy dip of only 25 dB.

Parts b, c and d of Figure 3.29 show that there is usually a significant drop in energy between vowels and following consonants. The average energy change between nonsonorant consonants and preceding vowels and between nasals and preceding vowels is between 25 dB and 30 dB. However, between semivowels and preceding vowels, the average energy changes are only 14 dB and 18 dB. If we remove postvocalic consonants which are followed by a sonorant consonant, such as the /ɾ/ in the word "harlequin," the average energy change increases to 17 dB and 22 dB, and 43% of the vowel-semivowel transitions involve an energy decrease of more than 25 dB.



(a)

(b)

Figure 3.28: Measurement procedure for natural energy taper in word-final vowels. (a) Wide band spectrogram of the word "bourgeois." (b) Energy 640 Hz to 2800 Hz.

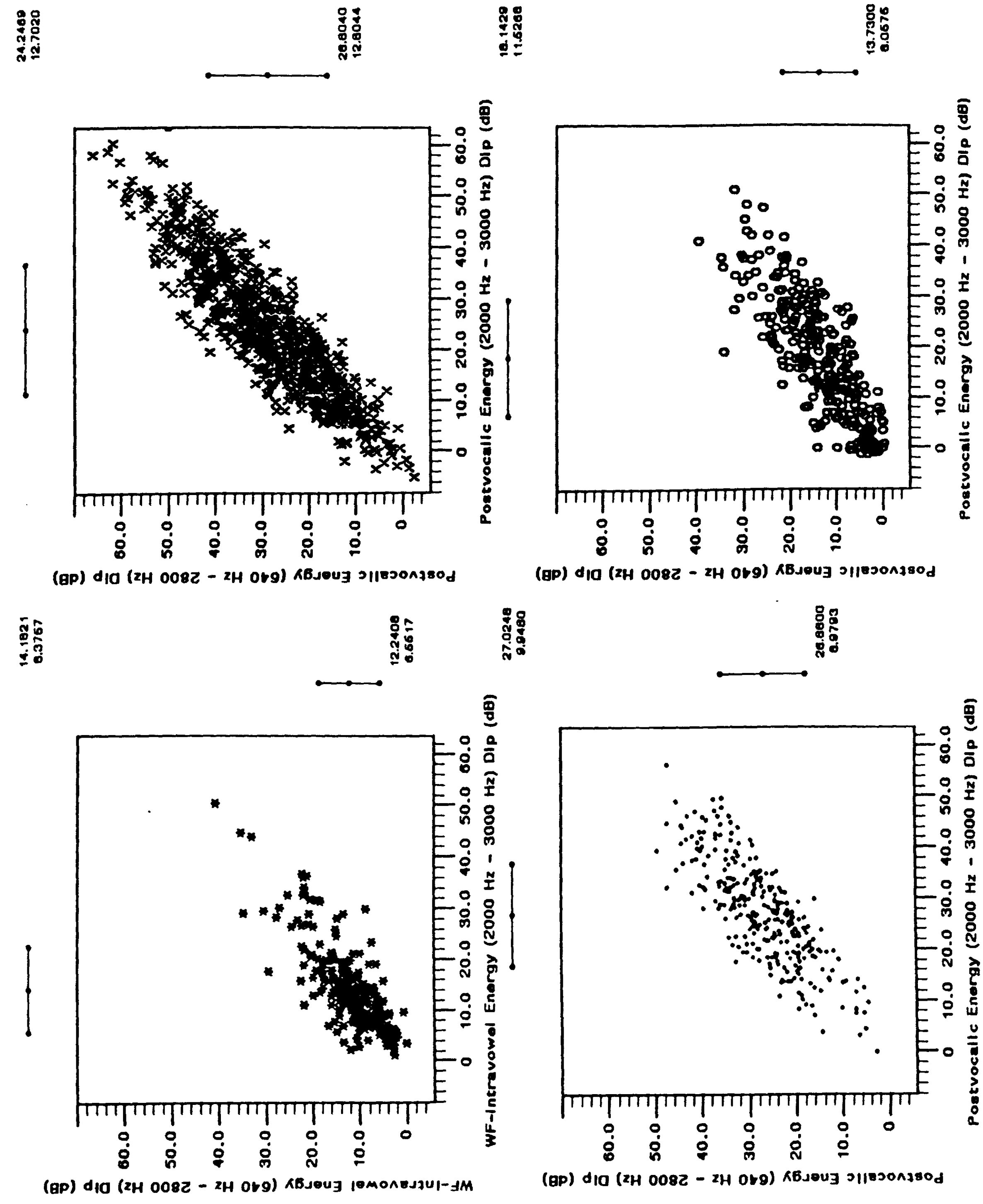
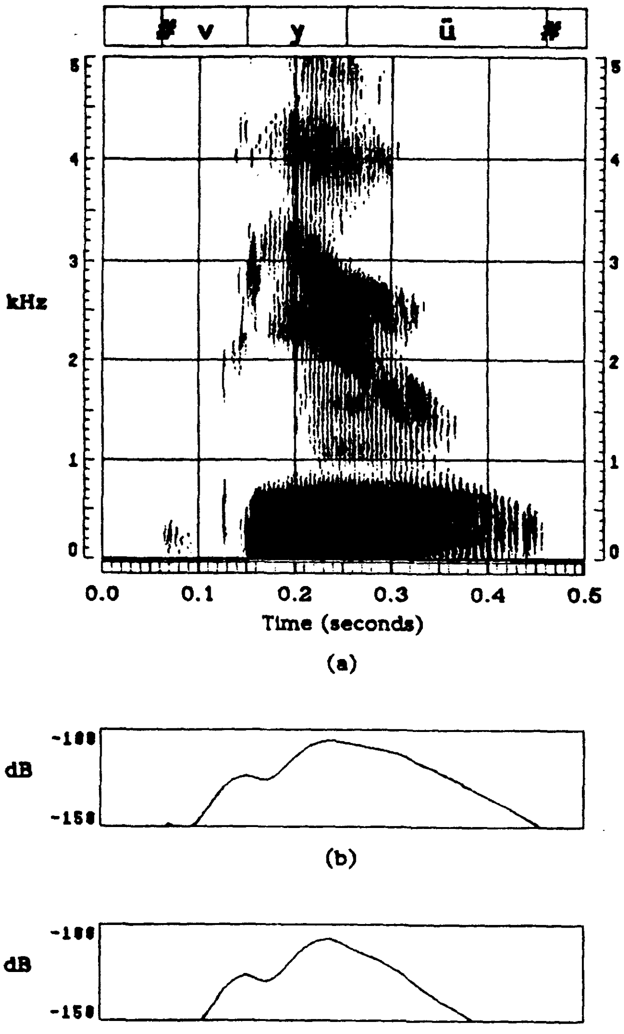


Figure 3.29: Comparisons of natural energy taper within vowels and average energy difference between postvocalic consonants and preceding vowels. vowels: *, nonsonorant consonants: x, nasals: ., semivowels: o.



(a)

(b)

(c)

Figure 3.30: Illustration of large energy taper in word-final diphthongs. (a) Wide band spectrogram of the word "view." (b) Energy 640 Hz to 2800 Hz.

3.2.5 Rate of Spectral Change

Fant (1960) observed that a distinguishing cue for /l/, when it precedes a vowel, is an abrupt shift in F1 from the /l/ into the following vowel. Dalston (1975) attributes this property to the rapid movement of the tongue tip away from the roof of the mouth. In addition, Dalston noted that this abrupt shift in F1 is often accompanied by a transient in the higher frequencies.

The parameter used in the study to extract this abrupt rate of change in energy between /l/ and vowels and, more generally, between consonants and vowels is based on the outputs of a bank of linear filters to which some nonlinearities (designed to model the hair-cell/synapse transduction process in the inner ear) are applied to enhance offsets and onsets (Seneff 1986). Compared to bandlimited energies based on the DFT, we found that these parameters have much sharper onsets and offsets. An example is shown in Figure 3.31 for the word "correlation." As can be seen, the abrupt spectral changes between /l/ and the surrounding vowels are captured in the waveforms, part b, which have sharp onsets and offsets between 300 Hz and 650 Hz and between 1070 Hz and 1700 Hz.

Based on these waveforms, we computed global onset and offset waveforms. The onset waveform is obtained by summing, in each frame, all the positive first differences in time (with a frame rate of 5 msec) of the channel outputs. Similarly, the offset waveform is computed by summing, in each frame, all the negative first differences in time. The resulting onset and offset waveforms for the word "correlation" are shown in parts c and d of Figure 3.31, respectively. As can be seen, the sharp spectral changes between the /l/ and the surrounding vowels show up in the onset and offset waveforms as a peak and a valley, respectively.

We examined the rate of change of these waveforms between all consonants and adjacent vowels. We defined the onset value to be the maximum rate of change between the consonant and following vowel. Likewise, we defined the offset value to be the maximum absolute value of the rate of change of the waveform between the preceding vowel and the consonant. As can be seen in Figure 3.31, the offset before the /l/ occurs at about 270 msec and the onset after the /l/ occurs at about 330 msec.

The data across all words and all speakers are discussed separately below for prevocalic, intervocalic and postvocalic consonants. In each context, we compare the rate of change associated with the semivowels with those associated with the nasals and nonsonorant consonants. In addition, we compare the rate of spectral change associated

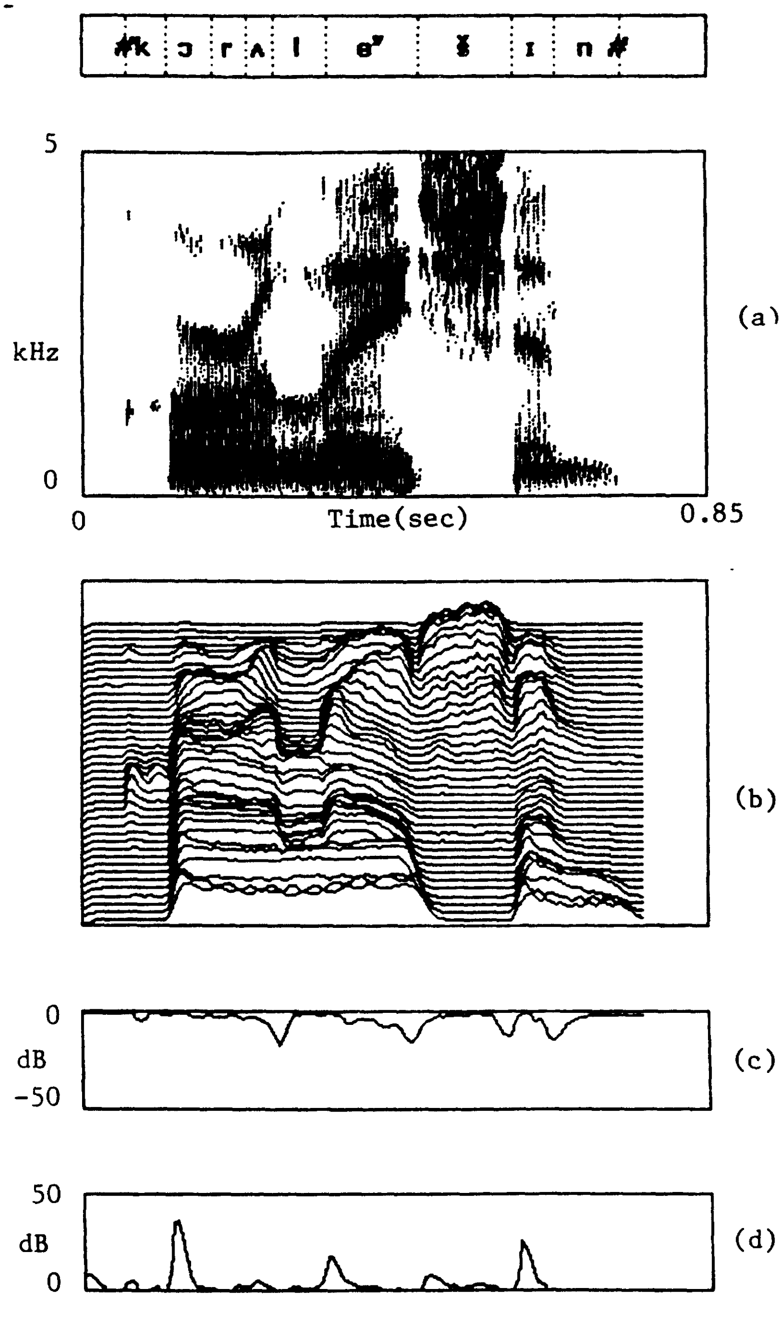


Figure 3.31: An illustration of parameters which capture abrupt spectral changes. (a) Wide band spectrogram of "correlation." (b) Channel outputs of an auditory model. (c) Offset waveform. (d) Onset waveform.

with /l/'s with those associated with the other semivowels.

Prevocalic Consonants

Only onsets are associated with prevocalic consonants since they are not preceded by vowels. Since the semivowels can be devoiced in this case, we only examined the onsets between semivowels which were either word-initial or preceded by a voiced consonant. These data, along with onset values associated with prevocalic nonsonorant consonants and nasals, are compared in Figure 3.32.

As expected, the average onset values associated with nonsonorant consonants and nasals, shown in parts a and b, are larger than those associated with the semivowels, shown in parts c and d. In addition, the average onset value associated with /l/, part c, is larger than that of the other semivowels, part d. However, as can be seen, there is a wide spread in the distribution of onset values. It appears as if stress is a major factor affecting the rate of spectral change between consonants and vowels. That is, the onset values tend to be large when the consonants precede vowels which are stressed, and small when the consonants precede vowels which are unstressed. Examples are shown in Figure 3.33. The onset value between the /l/ and /ʌ/ in "blurt" is 37 dB (at about 130 msec), whereas the onset value between the /l/ and /i^y/ in "linguistics" is only 5 dB (at about 155 msec). Similarly, small onset values between nasals and following vowels occur in words such as "misrule" and "misquote." An example of this phenomenon is also shown in Figure 3.33. In this case, the onset between the /m/ and /ɪ/ in "misrule" is only 2 dB (at about 110 msec).

Intervocalic Consonants

Since intervocalic consonants are surrounded by vowels, they have associated with them an offset and an onset. Figure 3.34 shows a comparison of the distribution of offset and onset values for intervocalic nonsonorant consonants, intervocalic nasals and intervocalic semivowels. The average and standard deviation of the offset and onset values appear with each scatter plot.

As in the prevocalic case, the average rate of spectral change associated with the nonsemivowel consonants, parts a and b, is greater than the average rate of spectral change associated with the semivowels, parts c and d. In addition, the average onset and offset values between /l/'s and surrounding vowels, part c, is greater than the ones between the other semivowels and adjacent vowels, part d. Again, stress appears to

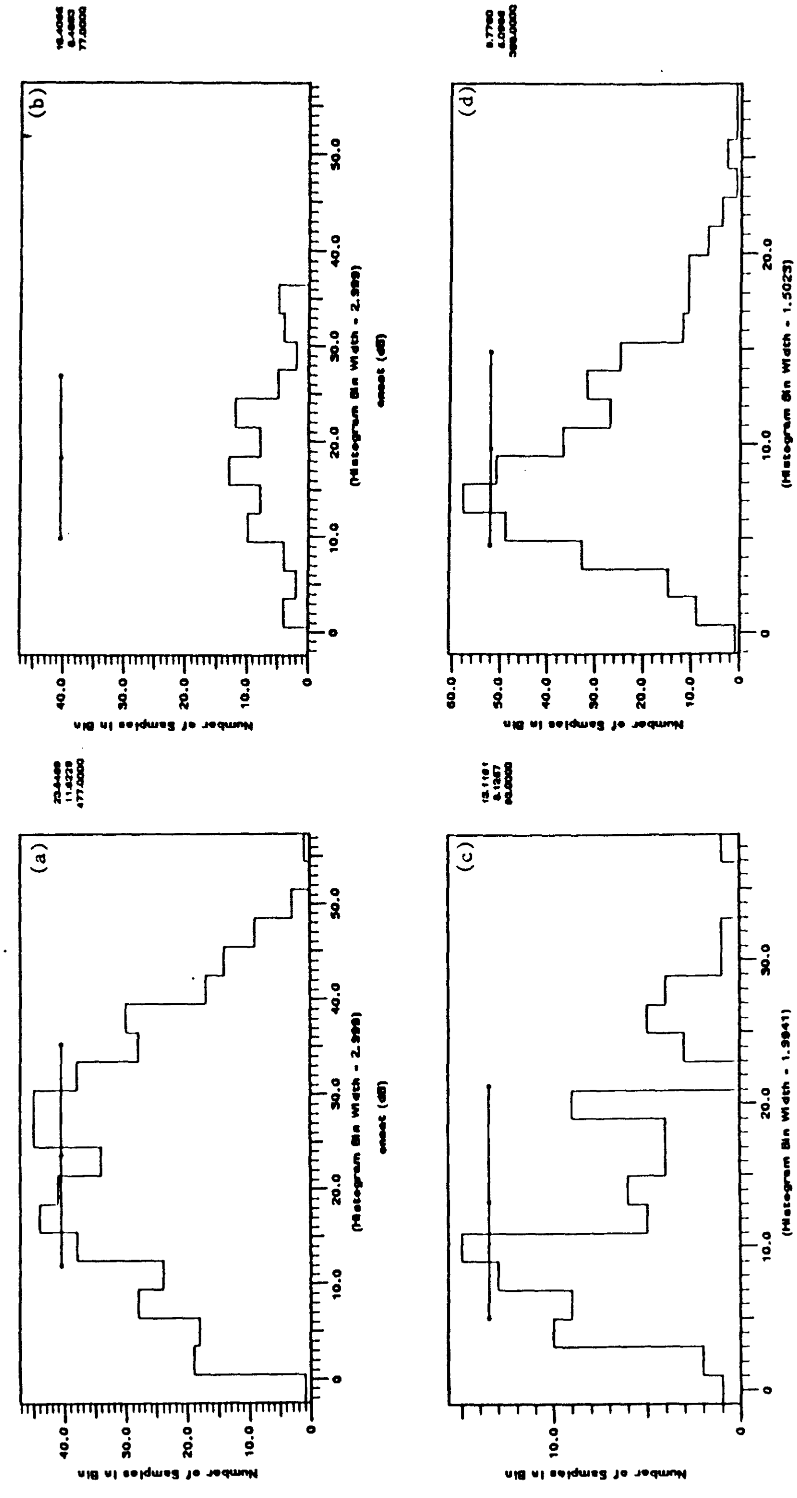


Figure 3.32: Onsets between following vowels and (a) prevocalic nonsonorant consonants (b) prevocalic nasals (c) /l/'s and (d) other semivowels.

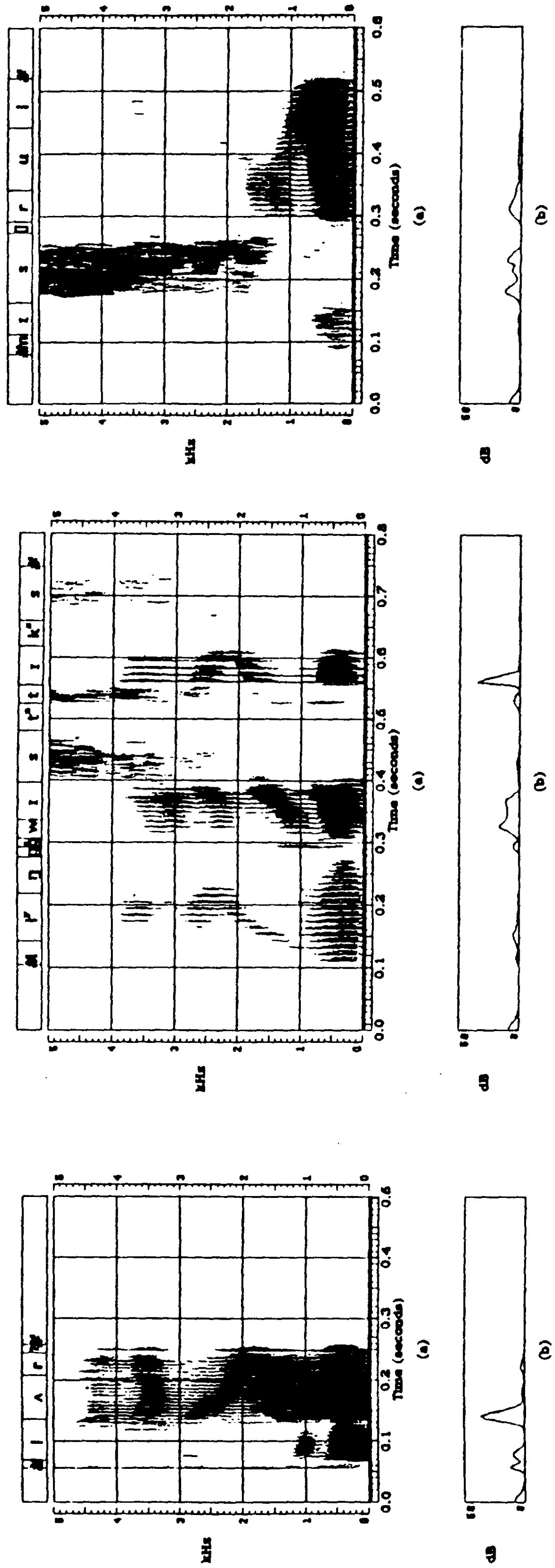


Figure 3.33: Rate of spectral change associated with prevocalic /l/'s in "blurt," "linguistics" and "misrule." (a) Wide band spectrograms. (b) Onset waveform.

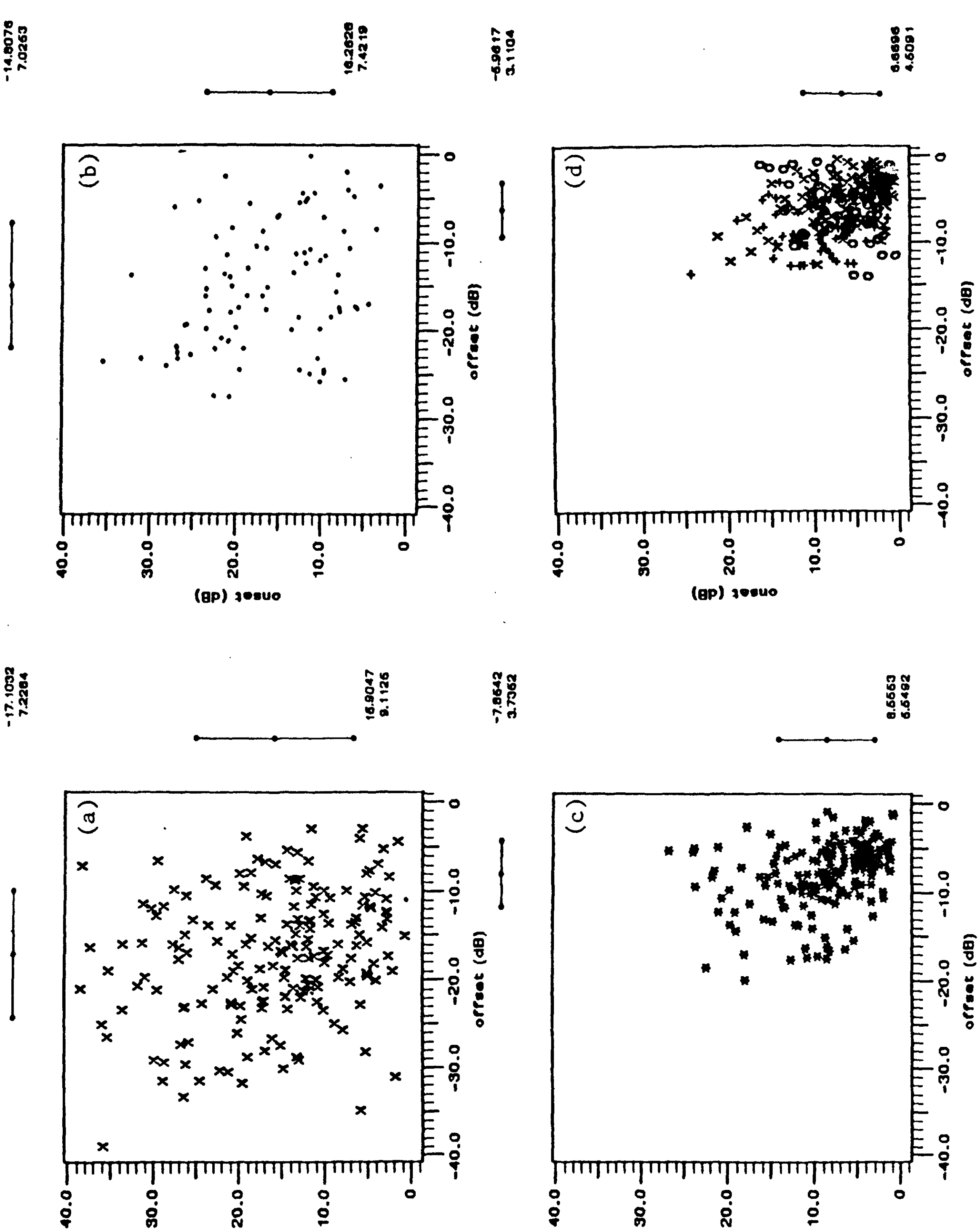


Figure 3.34: Onsets and Offsets between surrounding vowels and intervocalic (a) nonsonorant consonants (b) nasals (c) /l/'s and (d) other semivowels.

be a major factor affecting the rate of spectral change. That is, those /l/'s associated with the higher onset values occur before stressed vowels. Examples are the /l/'s in "roulette" and "caloric." Similarly, those /l/'s associated with offset values less than -15 dB also occur before stressed vowels, such as those in the words "poilu" and "walloon." In addition, some /l/'s with abrupt offsets occur before vowels which have secondary stress, such as those in "twilight" and "emasculate." Shown in Figure 3.35 is the word "walloon" which has an abrupt offset between the /l/ and the preceding vowel, and an abrupt onset between the /l/ and the following vowel. As can be seen, the offset before the /l/ occurs at -190 msec and is -18 dB. The onset after the /l/ occurs at 260 msec and is 22 dB.

As in the case of prevocalic /l/'s, some intervocalic /l/'s are associated with a gradual rate of spectral change. Such /l/'s usually occur after stressed vowels and before unstressed vowels, such as those in the words "swollen" and "horology," or they occur between unstressed vowels, such as the second /l/ in "soliloquize" and the intervocalic /l/ in "calculus." This latter result is not surprising given the data of Section 3.2.4, which show that intervocalic /l/'s in this context may not have significantly less mid-frequency energy than the surrounding vowels. For comparison, we included in Figure 3.35 the word "swollen," which has a gradual rate of spectral change between the /l/ and surrounding vowels. In this case, the offset is only -7 dB (at about 350 msec) and the onset is only 9.8 dB (at about 410 msec).

Postvocalic Consonants

The distribution of offset values associated with the postvocalic consonants are compared in Figure 3.36. As can be seen, the spread of offset values associated with the nonsemivowels, parts a and b, is much wider than the distributions associated with /l/ and /r/, parts c and d, respectively. Note that there is not a marked difference between the latter distributions. This result suggests that, in the case of postvocalic /l/'s, the tongue tip may not make contact with the palate. Or, it if does, it's release from the roof of the mouth is gradual.

3.2.6 Dip Region Duration

The data given in Sections 3.2.4 and 3.2.5 show that, when the semivowels occur intervocalically, they usually have less energy than both of the surrounding vowels, such that they have associated with them an offset and an onset. The offsets and

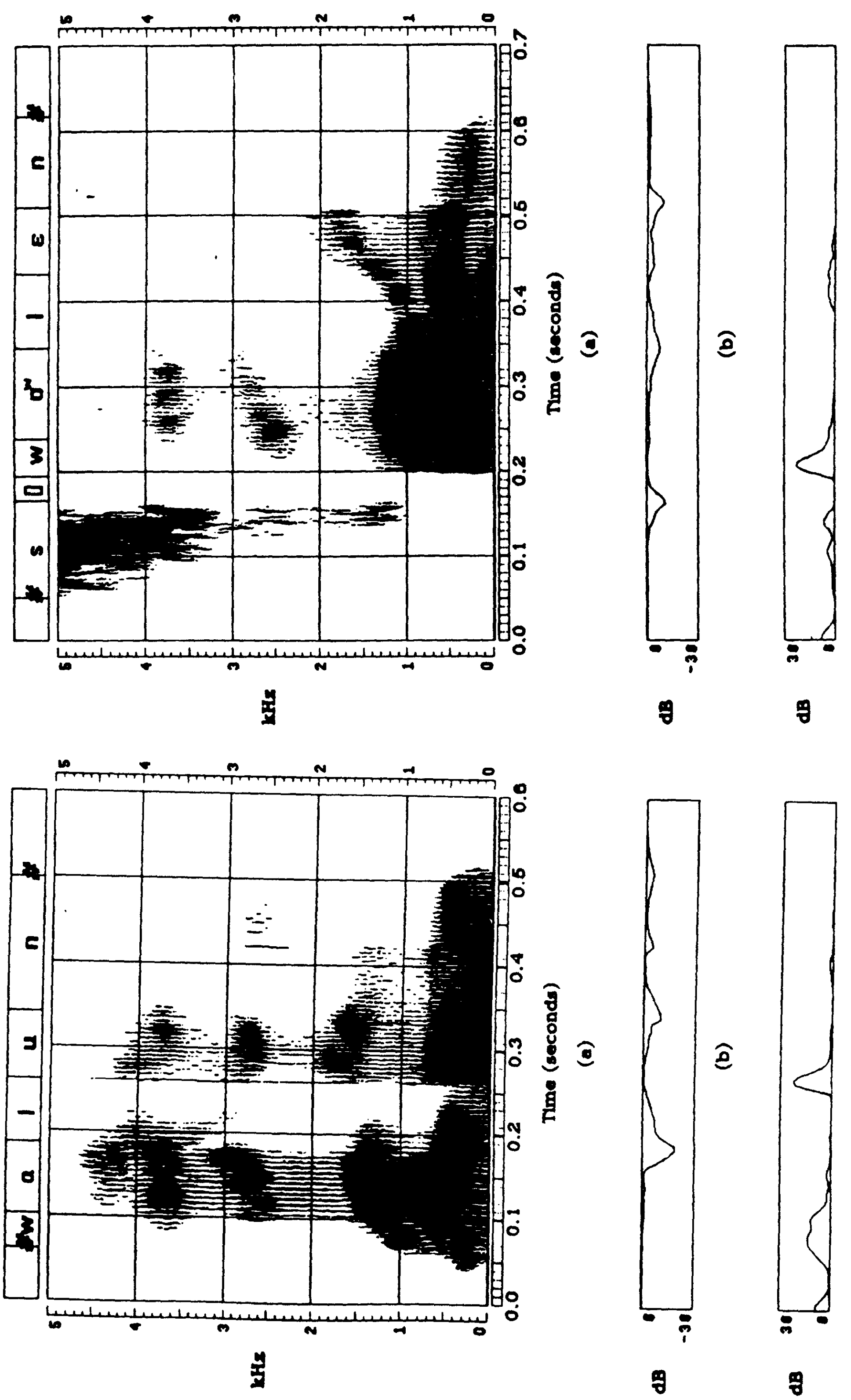


Figure 3.35: Rate of spectral change associated with intervocalic /l/'s in "walloon" and "swollen." (a) Wide band spectrograms. (b) Offset waveform. (c) Onset waveform.

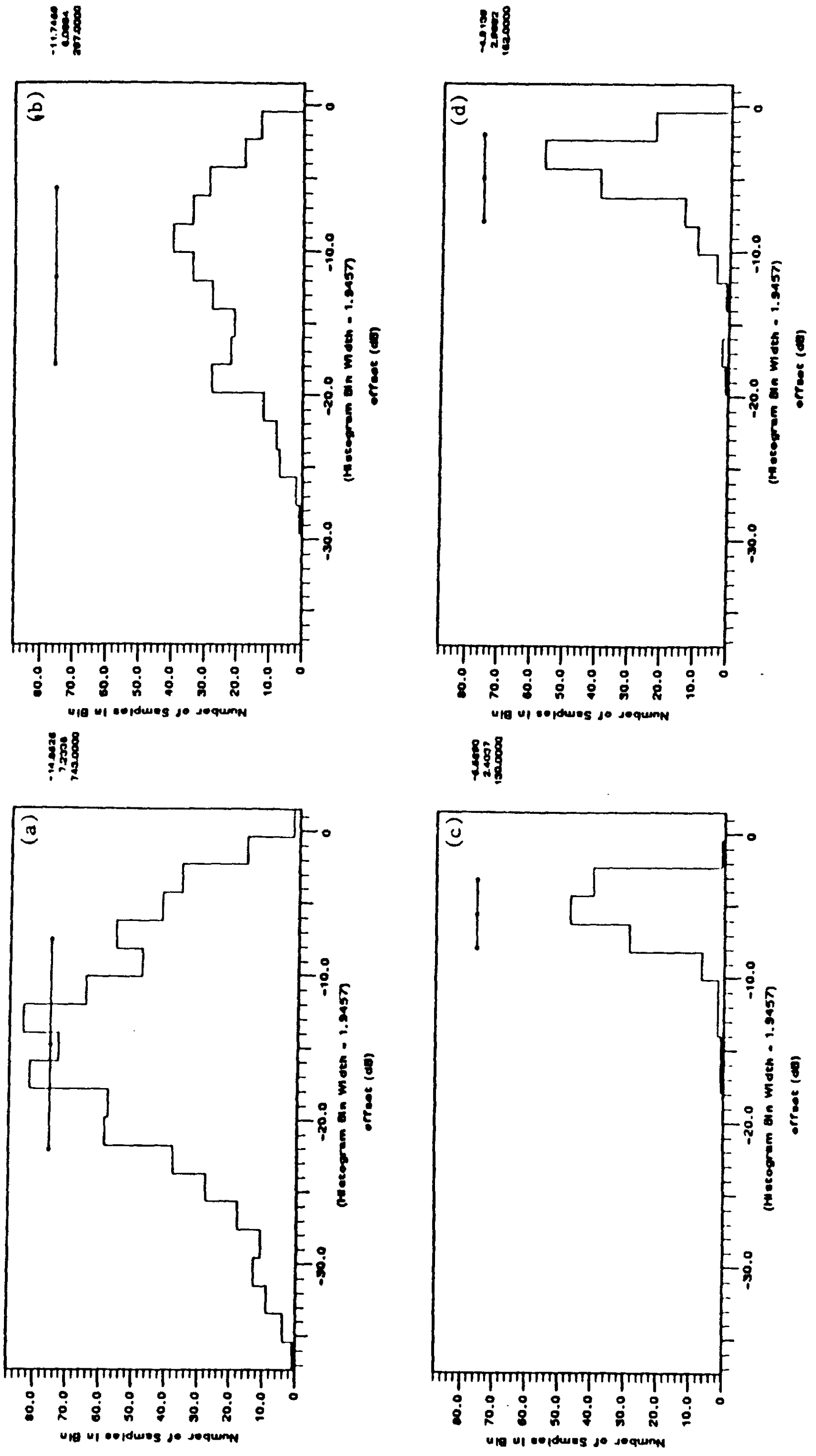


Figure 3.36: Offsets between preceding vowels and postvocalic (a) nonsonorant consonants (b) nasals (c) /l/'s and (d) other semivowels.

onsets can be considered to correspond to the beginning and end of the semivowels. We shall define the time difference between them to be the duration of the energy dip region. This correspondence can be seen in the word "correlation" shown in Figure 3.31 where the difference between the time of the offset occurring between the /l/ and the preceding vowel, and the time of the onset occurring between the /l/ and the following vowel, is equal to the duration of the intervocalic dip region.

In this part of our acoustic study, we compare the duration of the energy dip regions when there is either one or two sonorant consonants occurring between vowels. We have observed that when two sonorant consonants occur between vowels and the first consonant is a semivowel (in which case it has to be either an /l/ or /r/ since only they can be in postvocalic position), then the offset between the preceding vowel and the intervocalic sonorant consonant cluster usually occurs after the semivowel, at the beginning of the following sonorant consonant. This type of energy change is illustrated with the word "harmonize" shown on the left side of Figure 3.37. This word contains the intervocalic sonorant consonant cluster /rm/. As can be seen, the offset occurs after the /r/ at the beginning of the /m/, and the onset occurs at the boundary between the /m/ and the following vowel. Thus, only the /m/ is included in the energy dip region which is 75 msec in duration.

On the other hand, when the first member of an intervocalic sonorant consonant cluster is a nasal, then the energy offset will occur before this sonorant consonant. This type of energy change is illustrated with the word "unreality" shown on the right side of Figure 3.37. In this case, the intervocalic sonorant consonant cluster is /nr/. As can be seen, the offset between the sonorant consonant cluster and the preceding vowel occurs before the /n/ at about 175 msec, and the onset occurs after the /r/ before the following vowel at about 295 msec. Thus, the energy dip region includes both sonorant consonants and is 120 msec in duration.

Thus, by comparing the time difference between the offsets and onsets surrounding the intervocalic sonorant consonant clusters, we see that the duration of the energy dip region is usually much longer when the first member of the cluster is not a liquid, than when the first member of the cluster is a liquid.

The results of the difference in duration (measured in frames where the frame rate is 5 msec) between energy dip regions which contain only one intervocalic sonorant consonant (a semivowel or nasal), an intervocalic sonorant consonant cluster where the first member is a liquid, and an intervocalic sonorant consonant cluster where the first

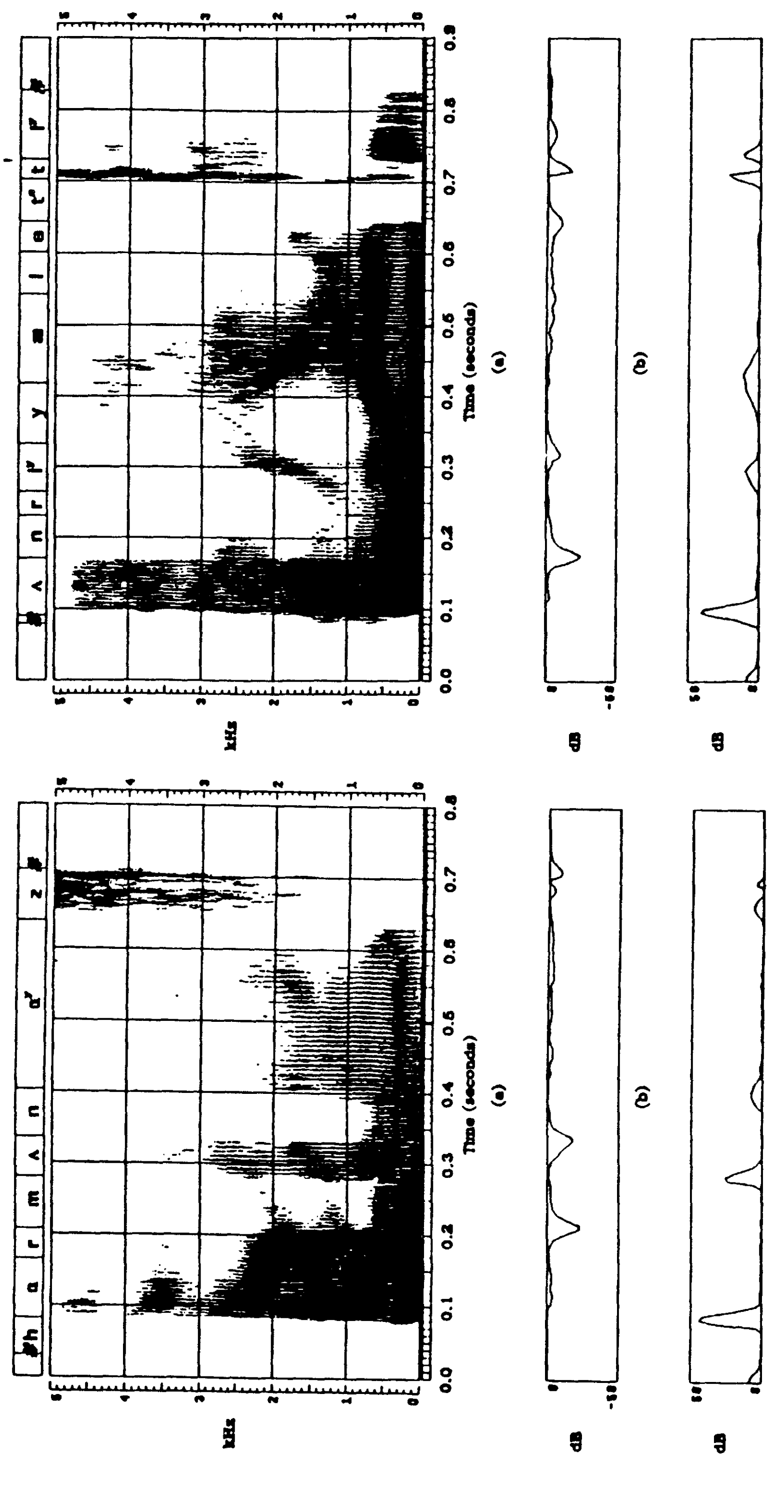


Figure 3.37: Comparison of the duration of the energy dip regions in "harmonize" (left) and in "unreality" (right). (a) Wide band spectrograms. (b) Offset waveforms. (c) Onset waveforms.

member is a nasal are compared in Figure 3.38. As can be seen, the average duration of energy dip regions containing only one sonorant consonant is comparable to the average duration of energy dip regions involving a sonorant consonant preceded by a liquid. This result suggests that the liquid is not included in the energy dip region.

On the other hand, energy dip regions involving a sonorant consonant which is preceded by a nasal are about 12 frames or 60 msec longer than energy dip regions containing only one sonorant consonant, and 10 frames or 50 msec longer than energy dip regions involving a sonorant consonant preceded by a liquid. This result suggests that this type of dip region contains both sonorant consonants. In fact, many of the latter dip regions with durations that overlap with the former cases are short because the nasal does not appear as a separate segment, but is manifested by nasalization within the vowel.

3.3 Discussion

This acoustic study is an evaluation of two factors. First, it is an assessment of the effectiveness of the selected parameters and measures used in capturing the desired acoustic properties. Clearly, in some cases, better attributes and more precise measures can be developed. For example, the grouping of some /k/'s, which have low-frequency bursts with nasals and semivowels on the basis of the properties used to extract the features *voiced* and *sonorant* (see Section 3.2.3) is undesirable. Second, this study is an analysis of how humans produce speech. For example, the inclusion of some voiced fricatives and stops with voiced and sonorant consonants appears to be reasonable. The data show that when these consonants occur between sonorant segments, there can be considerable feature assimilation, such that they look sonorant as well.

In addition, the results seem to suggest that some features are distinctive while others are redundant. For example, the data in Tables 3.6 - 3.8 (see pages 64 and 65 show that /r/ almost always has a lower F3 value than that of the adjacent segment(s). In the cases where this is not true, the vowel is r-colored with an F3 frequency at or below 2000 Hz. Thus, it appears that the feature ***retroflex*** is always present, although its acoustic correlate, due to feature assimilation, may have varying degrees of strength. On the other hand, the data of Section 3.2.4 show that 14% of the intervocalic /r/ segments are not significantly weaker than the surrounding vowels. That is, the /r/ does not always appear to be nonsyllabic. One interpretation of these results is that

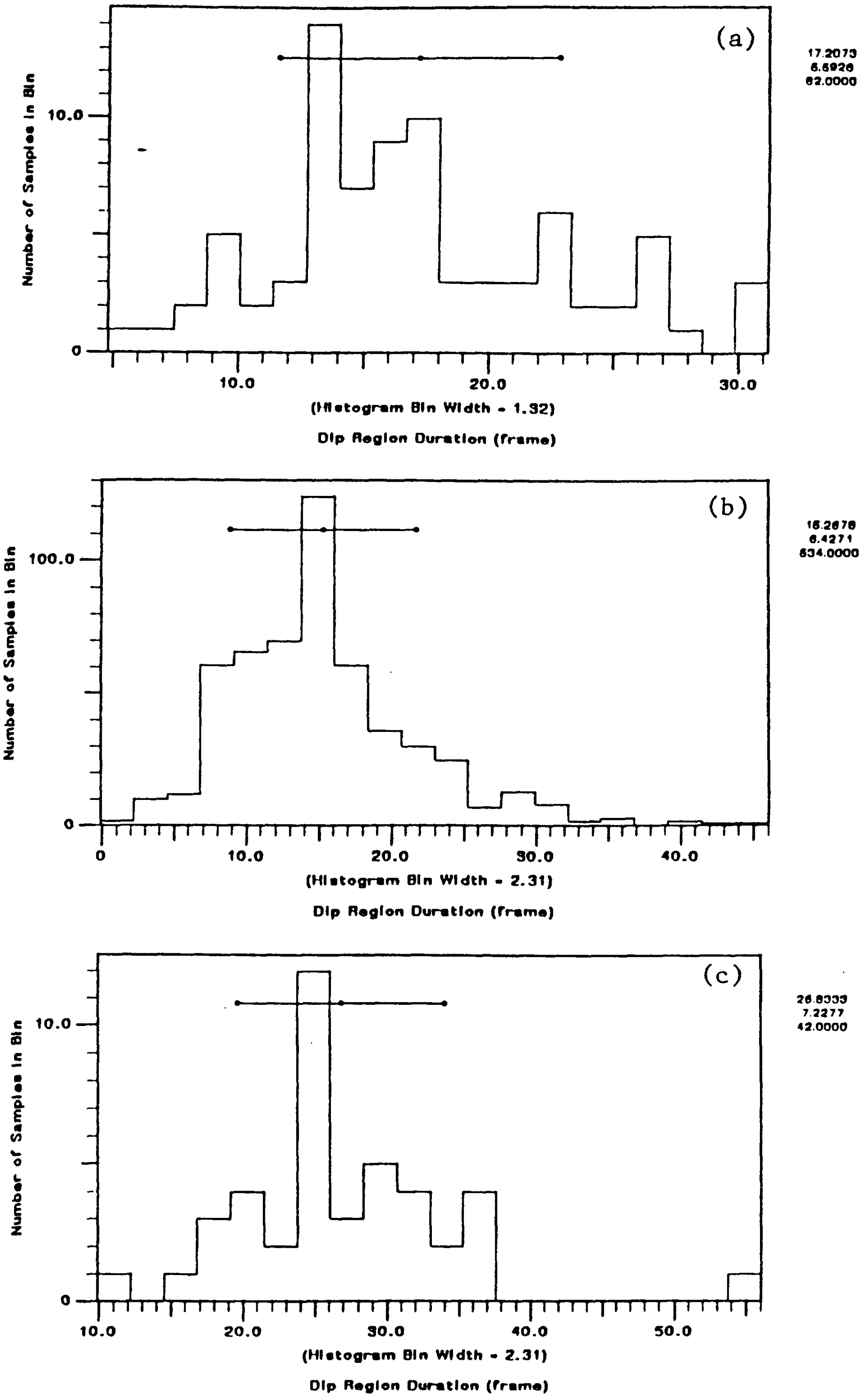


Figure 3.38: Durations of intervocalic energy dip regions containing (a) a semivowel or nasal (b) a sonorant consonant cluster when first member is a liquid and (c) a sonorant consonant cluster when first member is a nasal.

for /r/, the feature *retroflex* is distinctive, but the feature *nonsyllabic* is redundant.

The data provide further support for the theory of redundancy in speech (Stevens et al., 1986). While each of the properties investigated provides some separation between the desired sounds, there remains some overlap. No one property always provides a clear distinction. Instead, some discriminations require the integration of several acoustic cues. For example, the data in Figures 3.3, 3.4 and 3.5 show that there is some overlap between the /r/'s and other semivowels on the basis of F3-F0. That is, because of feature assimilation effects, F3 may not always be at a low enough frequency such that on the basis of it alone, we can determine that the segment is an /r/. In such cases, additional cues, such as the direction and extent of the transition of F3 between the /r/ and the adjacent sound(s) and the spacing between F3 and F2 within the /r/ segment, may be needed before the /r/ can be correctly identified. Though there are presently no features for which these additional cues are acoustic correlates, they do appear to be needed for recognition of /r/.

Several general tendencies have been observed in the data. First, an F2 minimum always occurs in a /w/ segment. This acoustic event enhances the detection of the feature *back*. Second, an F2 maximum always occurs in a /y/ segment. This acoustic event enhances the detection of the feature *front*. Similar tendencies occur for /l/ and /r/. That is, an F2 minimum and/or F3 maximum usually occurs in an /l/ segment and an F3 minimum usually occurs in an /r/ segment. However, due to feature assimilation, there are noteworthy exceptions.

In the case of /r/, an F3 minimum almost always occurs within its hand-transcribed region. However, as was discussed in Section 3.2.2, there are several exceptions to this pattern. The exceptions involve words like "cartwheel" and "harlequin," where the /r/ is followed by another consonant. In these cases, either an F3 minimum occurs in the vowel or F3 stays relatively constant at a low frequency throughout what can be called the vowel and /r/ region. That is, acoustically, the vowel and /r/ appear to be completely assimilated such that the resulting segment is an r-colored vowel. For example, consider the first sonorant regions in the four repetitions of the words "cartwheel" shown in Figure 3.39. As can be seen, F3 remains fairly constant at or slightly below 2000 Hz in each case. No discernible acoustic cue points to two separate /ɑ/ and /r/ segments.

These acoustic data provide evidence for the syllable structure as explained by Selkirk (1982, and others therein). This syllable structure is shown in Figure 3.40,

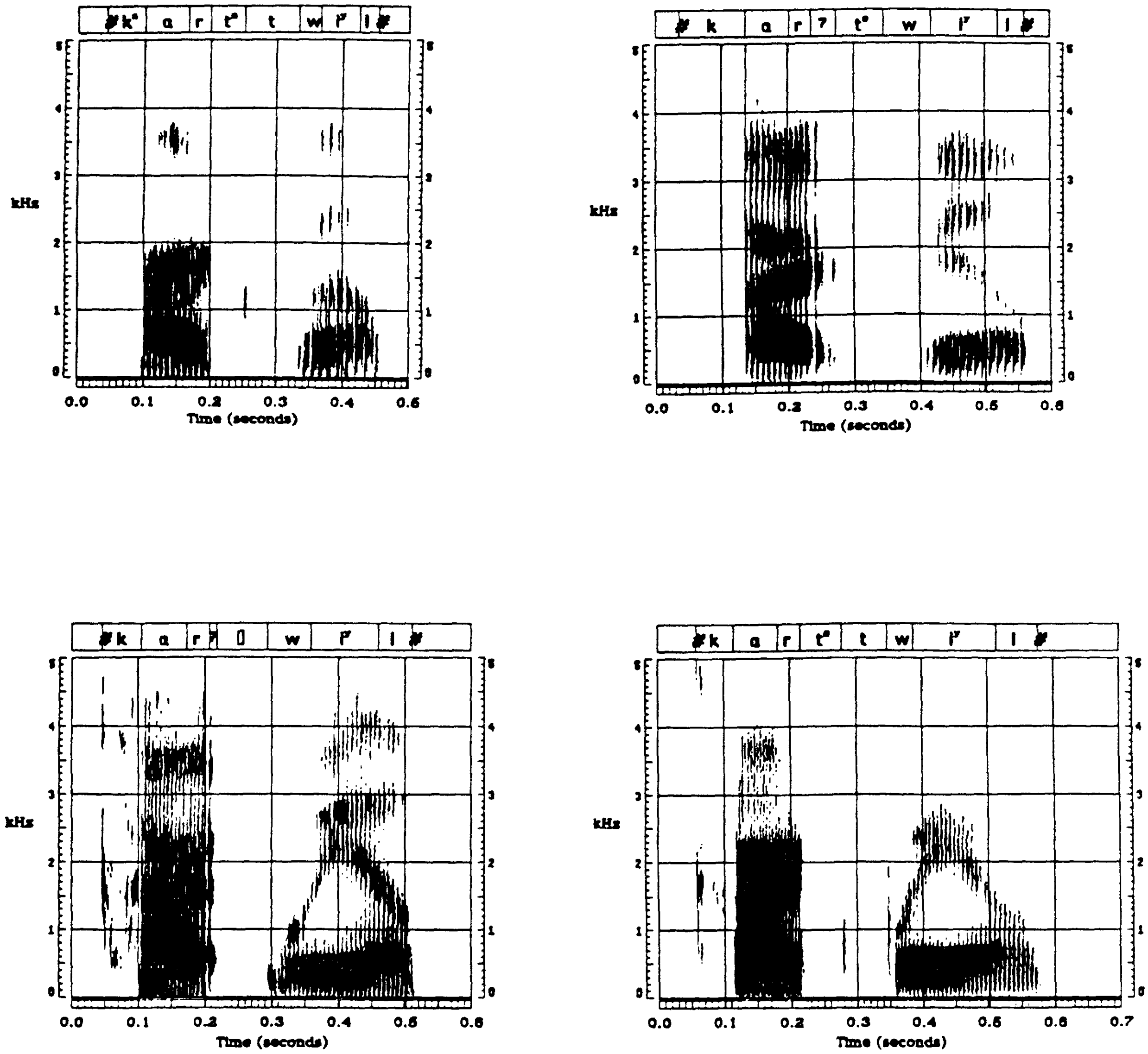


Figure 3.39: Wide band spectrograms of the word "cartwheel" spoken by each speaker. In each word, the /ɑ/ and /r/ sounds appear to be merged into one segment.

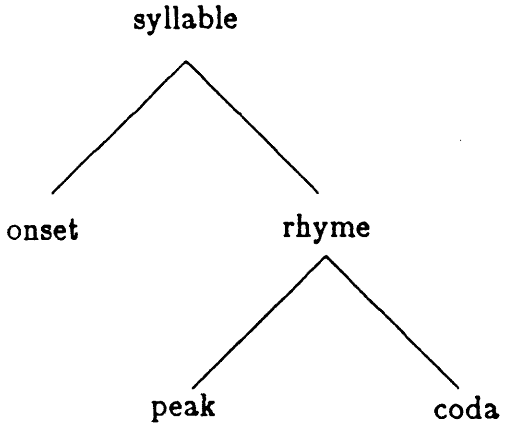


Figure 3.40: Tree structure for syllable.

where the onset consists of any syllable-initial consonant sequence, the peak consists of either a vowel or vowel and sonorant, and the coda consists of any syllable-final consonant sequence. Selkirk states that when a postvocalic liquid is followed by a consonant which must occupy the syllable-final position, the liquid will be part of the peak. Based on this theory, the structure for the first syllable in "cartwheel" is as shown in Figure 3.41. Thus, this theory accounts in a natural way for some overlap in the features of the vowel and liquid.

When postvocalic liquids are not followed by a consonant which must be syllable-final, Selkirk states that they tend to be consonantal though they have the option of being part of the peak or the coda. In the case of /r/, the acoustic data suggest that both situations occur. Compare the spectrograms of the words "harlequin," "carwash" and "Norwegian" shown in Figure 3.42. In the cases shown in the first row, the vowel and /r/ appear to be one segment in the sense that retroflexion extends over the entire vowel duration. Thus, it appears as if they are both a part of the syllable peak. On the other hand, in the cases shown in the second row, the vowels do not appear to be retroflexed. Instead, there is a clear downward movement in F3 which separates the vowel and /r/. Thus, in these cases, the /r/ is probably in the coda.

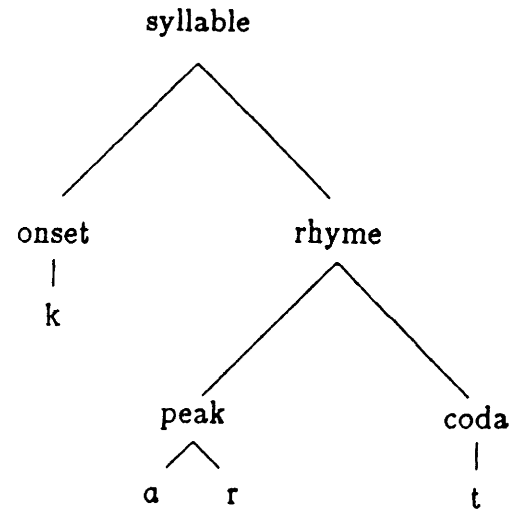


Figure 3.41: Tree structure for first syllable in "cartwheel."

Although a more extensive study is needed before any conclusive statements can be made regarding this phenomenon, it appears from these data that there should be no exception clause in the phonotactic constraints of semivowels for words like "snarl," where the /l/ is supposedly separated from the vowel by the /r/. Instead, it appears that the semivowels always occur adjacent to vowels, even in words like "snarl." In cases such as this, the vowel and /r/ probably both make up the syllable nucleus.

Spectrograms of the word "snarl" spoken by each speaker are shown in Figure 3.43. Even though they are not transcribed as such, the two occurrences of "snarl," shown in the top row, were pronounced as /snɑrəl/ with an intervocalic /r/. Consequently, there is a significant dip in F3. A /ə/ was not inserted between the /r/ and /l/ in the first occurrence on the bottom row. In this case, F3 remains constant at a low frequency, such that the vowel and /r/ appear to be completely assimilated. Finally, it is not clear whether the last occurrence was pronouned as /snɑrl/ or /snɑrəl/. Regardless of how it was pronounced, a steady F3 frequency at about 2100 Hz can be traced throughout most of the vocalic region.

Further support for this type of feature assimilation was given in Section 3.2.6, where the data show that postvocalic liquids that are in an intervocalic sonorant

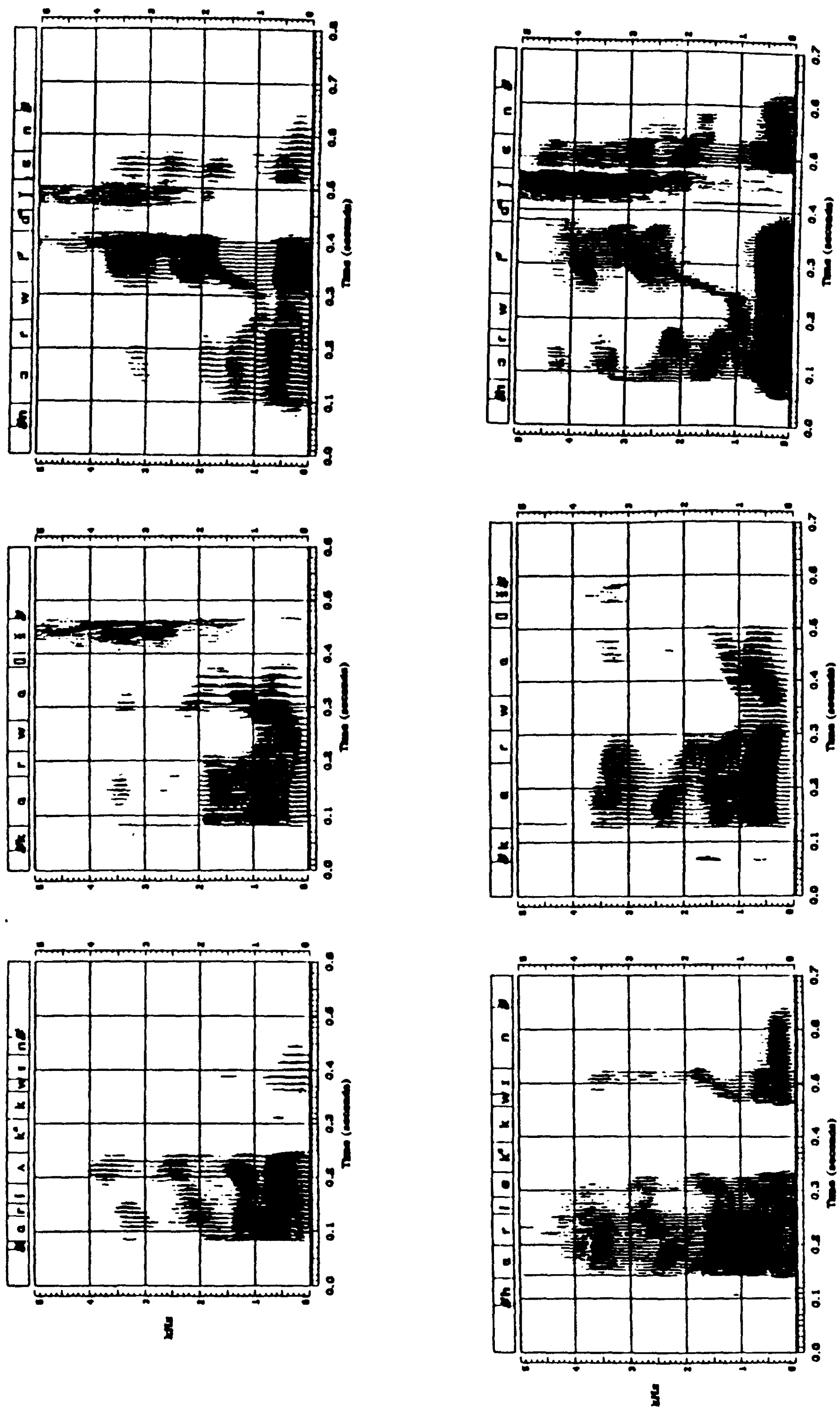


Figure 3.42: Wide band spectrograms of the words "harlequin," "carwash" and "Norwegian," each spoken by two different speakers. In each word in the top row, the /r/ and preceding vowel appear to be merged into one segment. In each word in the bottom row, the /r/ and preceding vowel appear to be separate segments.

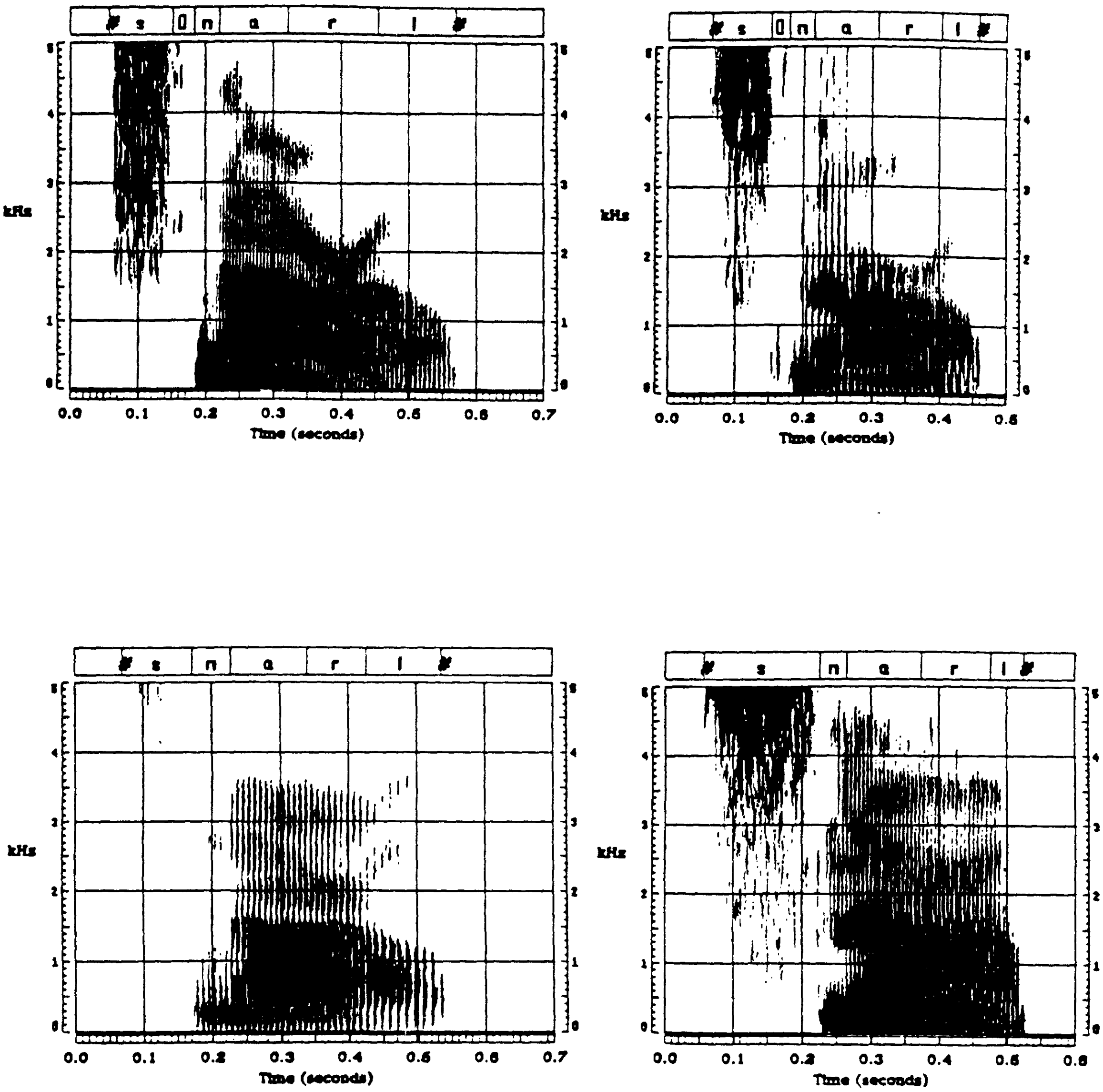


Figure 3.43: Wide band spectrograms of the word "snarl" spoken by each speaker. The /l/ is either adjacent to a /ə/ which has been inserted between the /r/ and /l/, or it is adjacent to a retroflexed vowel.

consonant cluster are not part of the energy dip region. Therefore, on the basis of significant energy change, they do not appear to be nonsyllabic. Although this result would seem to suggest that all such postvocalic liquids are part of the syllable nucleus, we feel that in some cases there may be other cues which signal their consonantal status. A case in point are the significant F3 dips occurring within the /r/'s in the second row of words in Figure 3.42. It appears as if this acoustic event is used to separate the /r/ from the vowel.

This point brings us to our final discussion of the semivowel /l/. The data in Tables 3.6 - 3.8 show that an F2 minimum usually occurs within an /l/ segment. Furthermore, the data show that an F3 maximum also occurs within many of the /l/ segments, particularly in the postvocalic allophones. However, much of the discussion for /r/ applies for /l/ as well. That is, it appears as if postvocalic, but not word-final, /l/'s will sometimes be part of the syllable nucleus and sometimes part of the coda. In words like "bulrush," "walnut" and "almost," a clear /l/ is not always heard. Many times, no discernible acoustic cue separates the /l/ from the preceding vowel. A case in point is the underlying /l/ in the word "almost" shown at the top of Figure 3.44. As can be seen, an /l/ was not included in the transcription of this word. However, in some repetitions of these words, there is a significant rise in F3 before the energy dip region. This acoustic event could be the cue which signals a separate /l/ segment. An example of this phenomenon is shown at the bottom of Figure 3.44, where a spectrogram of the word "stalwart" is given. As can be seen, F3 rises about 200 Hz between the beginning of the /ɑ/ and the end of the /l/.

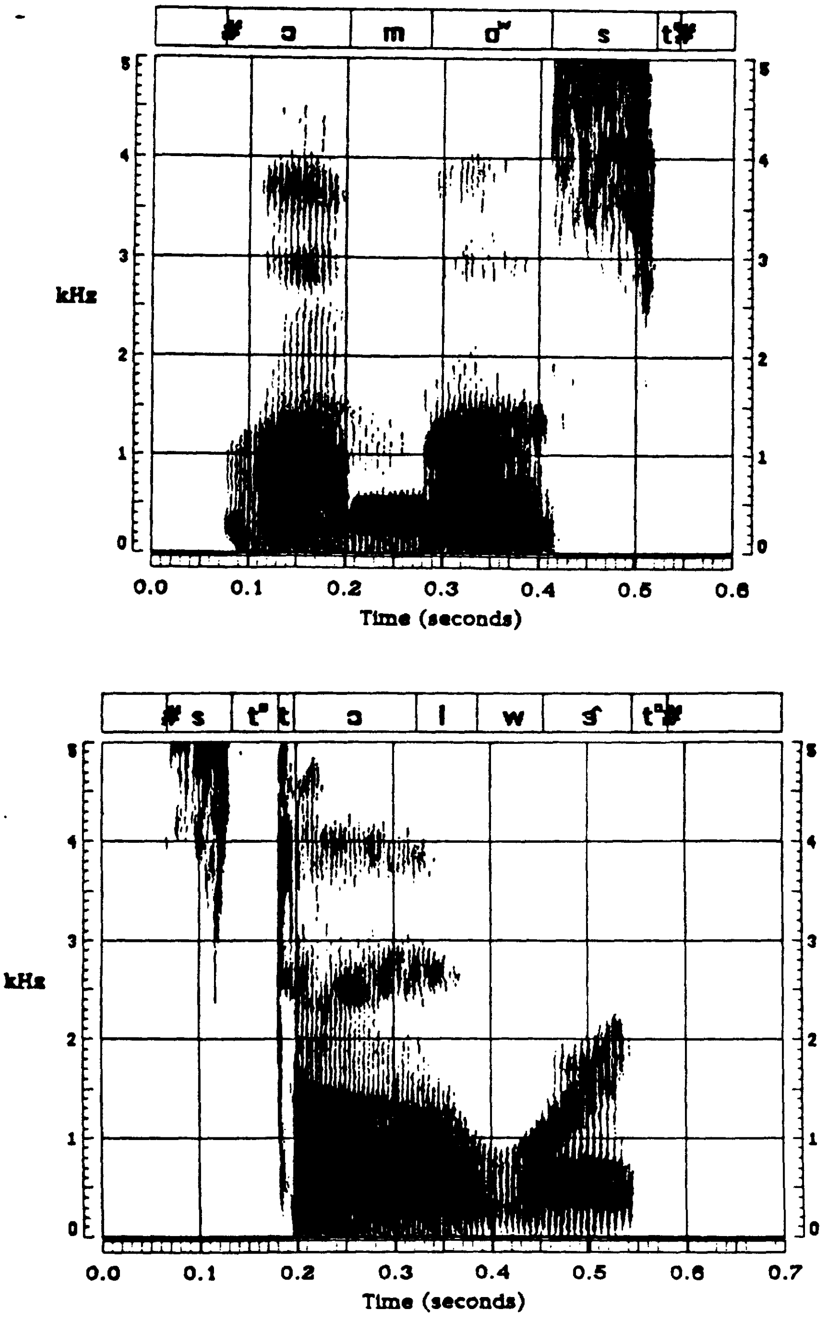


Figure 3.44: Wide band spectrograms of the words "almost" and "stalwart." In "almost," there does not appear to be a separate /l/ segment. In "stalwart," the rise in F3 from the beginning of the /ɔ/ is indicative of a separate /l/ segment.

Chapter 4

Recognition System

This chapter describes the recognition system in detail. The recognition process consists of four stages. First, the features needed to recognize the semivowels are specified. Second, these features are mapped into properties which are quantified. Third, algorithms are applied to automatically extract the properties. Finally, the properties are combined for recognition. Each stage is discussed in detail below.

4.1 Feature Specification

To recognize the semivowels, features are needed for separating the semivowels as a class from other sounds and for distinguishing among the semivowels. Shown in Tables 4.1 and 4.2 are the features needed to make these classifications. The features listed are modifications of those proposed by Jakobson, Fant and Halle (1952) and by Chomsky and Halle (1968). In the tables, a "+" means that the speech sound(s) indicated has the designated feature and a "−" means the speech sound(s) does not have the designated feature. If there is no entry, then the feature is not distinctive: For example, the data of Section 3.2.2 show that /l/ (except when it is postvocalic) and /r/ do not, in general, have as low an F2 frequency as /w/. In fact, Figure 3.2 shows that the difference between F2 and F1 of these semivowels can be as high as 1300 Hz. For this reason, the feature *back* in Table 4.2 is left unspecified for /r/ and prevocalic /l/'s.

This raises the question of why, in Table 4.2, we divided /l/ on the basis of whether it is prevocalic or postvocalic. This was done because of two distinct acoustic differences we observed between these allophones. As has been mentioned before, postvo-

Table 4.1: Features which characterize various classes of consonants

	voiced	sonorant	nonsyllabic	nasal
voiced fricatives, stops, affricates	+	−	+	−
unvoiced fricatives,stops,affricates	−	−	+	−
semivowels	+	+	+	−
nasals	+	+	+	+
vowels	+	+	−	−

Table 4.2: Features for discriminating among the semivowels

	stop	high	back	front	labial	retroflex
/w/	−	+	+	−	+	−
/y/	−	+	−	+	−	−
/r/	−	−		−	−	+
prevocalic /l/	+	−		−	−	−
postvocalic /l/	−	−	+	−	−	−

calic /l/'s generally have a closer spacing between F2 and F1 (average difference of 433 Hz) than prevocalic /l/'s (average difference of 693 Hz). In fact, the former difference is comparable to the average values obtained for prevocalic and intervocalic /w/'s (388 Hz and 422 Hz, respectively). For this reason, postvocalic /l/'s are considered to be *back*. In addition, the data of Section 3.2.5 show that the rate of spectral change (a first difference computed with a frame rate of 5 msec) is generally higher between prevocalic /l/'s and following vowels (13 dB) than between postvocalic /l/'s and preceding vowels (5.5 dB). This difference is even more pronounced when the adjacent vowels are stressed. In this case, abrupt spectral changes as high as 37 dB were observed between prevocalic /l/'s and following vowels. As stated earlier, this stop-like characteristic of /l/'s in this context is probably due to the rapid release of the tongue tip from the roof of the mouth in the production of this noncontinuant sound. In the case of postvocalic /l/'s, the tongue tip may never make contact with the roof of the mouth and, if it does, it's release is usually more gradual.

Unfortunately, since the transcriptions of the words do not include stress markers, we are unable to divide the intervocalic /l/'s into those which tend to be syllable-initial and those which tend to be syllable-final. However, we suspect, on the basis of the data presented in Section 3.2.5, that the intervocalic /l/'s which are syllable-initial tend to have abrupt offsets and abrupt onsets. Thus, in this sense, they resemble the prevocalic /l/'s. On the other hand, /l/'s which are syllable-final tend to have gradual offsets and gradual onsets. In this respect, they resemble the postvocalic /l/'s. Thus, the intervocalic /l/'s are assumed to be covered acoustically by the prevocalic and postvocalic /l/ allophones.

The feature specifications given in Tables 4.1 and 4.2 are based on canonic acoustic representations of the different speech sounds. However, as was shown in Chapter 3, the overlapping of features between adjacent phonetic segments can alter significantly their acoustic manifestation. As a result, the class and phonetic distinctions given in the tables cannot always be clearly made. For example, the results of Section 3.2.3 show that, in addition to the semivowels and nasals, other intersonorant voiced consonants sometimes exhibit the property of sonorancy.

Table 4.3: Mapping of Features into Acoustic Properties

Feature	Acoustic Correlate	Parameter	Property
Voiced	Low Frequency Periodicity	Energy 200-700 Hz	High*
Sonorant	Comparable Low & High Frequency Energy	Energy Ratio $\frac{(0-300)}{(3700-7000)}$	High
Nonsyllabic	Dip in Energy	Energy 640-2800 Hz	Low*
		Energy 2000-3000 Hz	Low*
Stop	Abrupt Spectral Change	Onset Waveform**	High
		Offset Waveform**	High
High	Low F1 Frequency	F1 – F0	Low
Back	Low F2 Frequency	F2 – F1	Low
Front	High F2 Frequency	F2 – F1	High
Labial	Downward Transitions for F2 and F3	F3 – F0	Low*
		F2 – F0	Low*
Retroflex	Low F3 Frequency & Close F2 and F3	F3 – F0	Low
		F3 – F2	Low

*Relative to a maximum value within the utterance.
**For a definition of these parameters, see Section 3.2.5.

4.2 Acoustic Correlates of Features

This section is divided into two parts. First, we will discuss the mapping of the features specified in Section 4.1 into measurable acoustic properties. This will be followed by a discussion of how the acoustic properties were quantified.

4.2.1 Mapping of Features into Acoustic Properties

Table 4.3 contains acoustic correlates of the features specified in Tables 4.1 and 4.2, the mapping of these features into properties which can be quantified and the parameters from which the properties are extracted. Note that there is no parameter from which we extract the acoustic correlate of the feature *nasal.* Thus, on the basis of Table 4.1, we expect the system to make some confusions between nasals and semivowels since they are both sonorant consonants.

The effectiveness of these properties in capturing the designated features was demonstrated in Chapter 3. Recall that the properties extracted from these parameters are based on relative measures which tend to make them insensitive to interspeaker and intraspeaker differences. The properties are of two types. First, there are properties which examine an attribute in one speech frame relative to another speech frame. For example, the property used to capture the *nonsyllabic* feature looks for a drop in either of two mid-frequency energies with respect to surrounding energy maxima. Second, there are properties which, within a given speech frame, examine one part of the spectrum in relation to another. For example, the property used to capture the features *front* and *back* measures the difference between F2 and F1. Some properties, such as the one which extracts the feature *sonorant*, keep nearly the same strength over intervals of time and, therefore, define regions within the speech signal. Other properties, such as that used to capture the feature nonsyllabic, are highlighted by maximum values of strength and, therefore, are associated with particular instants of time.

Based on our present knowledge of acoustic phonetics, some parameters, and therefore some properties, are more easily computed than others. For example, the different energy measures involve straightforward computations so that the energy-based properties are easily extracted. On the other hand, computation of the formant tracks is often complicated by nasalization and peak merging effects (see Section 2.2.3). Thus, the extraction of formant-based properties is not as reliable. Likewise, we have observed that the pitch tracks (Gold and Rabiner, 1969) are error prone at the beginning of voiced regions. For several frames in the beginning of a voiced region, the pitch frequency is sometimes registered as being several octaves higher than the average value within the utterance, or it is sometimes zero due to a considerable delay between the onset of voicing and the detection of periodicity by the pitch tracker. For this reason, the detection of voiced regions was based mainly on low frequency energy. However, pitch information was used to refine initial estimates.

4.2.2 Quantification of Properties

To quantify the properties, we used a framework motivated by fuzzy set theory (DeMori, 1983) which assigns a value in the range [0,1]. A value of 1 means we are confident that the property is present. Conversely, a value of 0 means we are confident that the acoustic property is absent. Values in between these extremes represent a fuzzy area with the value indicating our level of certainty that the property is present/absent.

As an example of how this framework is applied, consider the quantification of the acoustic property used to extract the feature ***nonsyllabic***. As discussed in Section 3.2.4, the acoustic correlate of this feature is significantly less energy in the consonant regions than in the vowel regions. In an attempt to define this property of "less energy" more precisely, we selected the bandlimited energies 640 Hz to 2800 Hz and 2000 Hz to 3000 Hz and examined their effectiveness in identifying the presence of intervocalic semivowels. Scatter plots comparing the range of values of the energy dips for vowels and intervocalic consonants are shown in Figure 3.22. Recall that less than 1% of the vowels contain an energy dip. Furthermore, these energy dips tend to be less than 2 dB.

Based on these data, this property was quantified into the regions shown in Figure 4.1. An energy dip of 2 dB or more definitely indicates a nonsyllabic segment. If an energy dip between 1 dB and 2 dB is measured, we are uncertain as to whether a nonsyllabic segment is present or not. Finally, energy dips of less than 1 dB are not indicative of a nonsyllabic segment.

Not all of the properties have a defined "maybe" region. Instead, "fuzziness" is expressed in slanted tails as opposed to abrupt cutoffs which would result in quantization. For example, consider the quantification of the property used to capture the features *back* and *front*. This property measures the difference between the first and second formants. Shown in part a of Figure 4.2 are overlays of smoothed distributions of F2-F1 for each of the semivowels. Based on this plot, we quantified this property into the four regions shown in Figure 4.2 b: very back, back, mid and front. Thus, a sound with an F2-F1 difference less than 300 Hz will be classified as very back with a confidence of 1, whereas a sound with an F2-F1 difference of 1500 Hz or more will be classified as front with a confidence of 1. On the other hand, a sound with an F2-F1 difference of 1450 Hz will be classified as front and mid with a confidence of 0.5

A listing of the qualitative descriptions given to the regions of the quantified properties is given in Table 4.4. As can be seen from this table, the number of regions within the quantified properties is variable. This number was based on the data as well as the type of discriminations needed to distinguish between the semivowels.

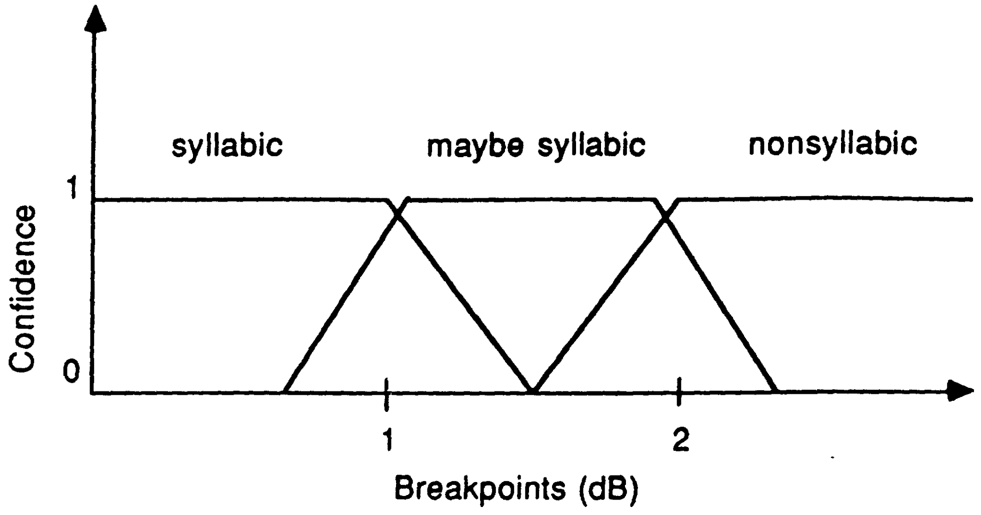


Figure 4.1: Quantification of the acoustic correlate of the feature *nonsyllabic.*

Table 4.4: Qualitative Description of Quantified Properties

Feature	Quantified Regions
Nonsyllabic	syllabic, maybe syllabic, nonsyllabic
Stop	gradual, abrupt, very abrupt onsets/offsets
High(Low)	high, maybe high, nonhigh, low, very low
Back(Front)	very back, back, mid, front
Retroflex	retroflex, maybe retroflex, not retroflex
	close f2 f3, maybe close f2 f3, not close f2 f3

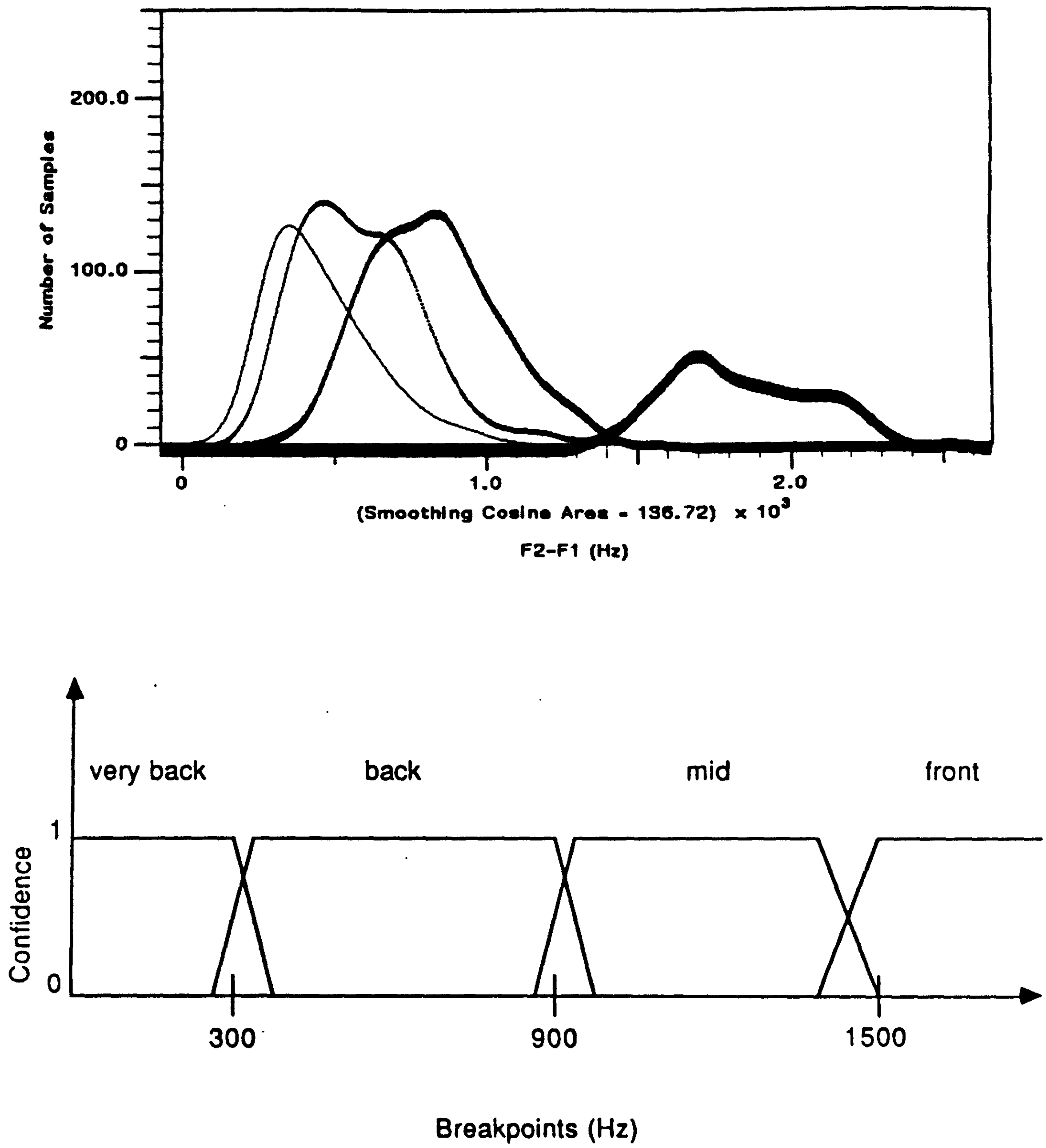


Figure 4.2: Quantification of the acoustic correlates of the features *back* and *front.*

4.3 Control Strategy

The recognition strategy for the semivowels is divided into two steps: detection and classification. The detection process marks certain acoustic events in the vicinity of times where there is a potential influence of a semivowel. In particular, we look for minima in the mid-frequency energies and we look for minima and maxima in the tracks of F2 and F3. Such events should correspond to some of the features listed in Tables 4.1 and 4.2. For example, an F2 minimum indicates a sound which is more "back" than an adjacent segment(s). Thus, this acoustic event will occur within most /w/'s and within some /l/'s and /r/'s. Note that acoustic events occurring within other sounds may be marked as well. For example, in addition to the semivowels, nasals and other consonants will usually contain an energy dip. Once all acoustic events have been marked, the classification process integrates them, extracts the needed acoustic properties, and through explicit semivowel rules decides whether the detected sound is a semivowel and, if so, which semivowel it is. At this time, by combining all the relevant acoustic cues, the semivowels should be correctly recognized while the remaining detected sounds should be left unclassified. A more detailed description of the recognition stages is given in this section.

4.3.1 Detection

The aim of this part of the recognition process is to mark all regions within an utterance where semivowels occur. To do this we use phonotactic constraints which restrict where the semivowels can occur within an utterance and, more specifically, within a voiced sonorant region. These constraints state that semivowels almost always occur adjacent to a vowel (with the exception of /rl/ clusters in words like "snarl"). Therefore, they are usually prevocalic, intervocalic or postvocalic. While all of the semivowels can occur in prevocalic and intervocalic positions, only the liquids /l/ and /r/ can occur in postvocalic positions.

These contexts map into three types of places within a voiced sonorant region. This mapping is illustrated in Figure 4.3. First the semivowels can be at the beginning of a voiced sonorant region. Semivowels of this type are prevocalic and they may be word-initial or in a cluster with a nonsonorant consonant(s). Second, the semivowels can be at the end of a voiced sonorant region. Semivowels of this type are postvocalic and they may be word-final or in a cluster with a nonsonorant consonant(s). Finally,

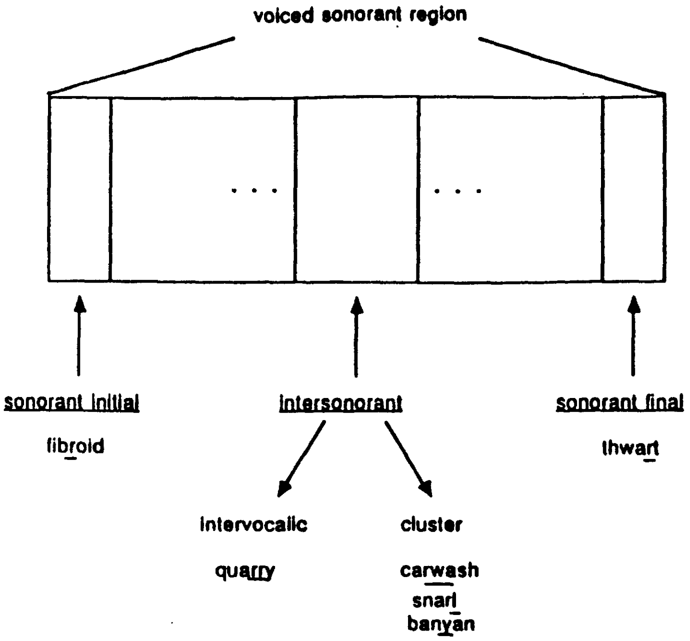


Figure 4.3: Places within a voiced sonorant region where semivowels occur.

the semivowels may be further inside a voiced sonorant region. We refer to these semivowels as intersonorant and one or more may be present. Semivowels of this type can be either intervocalic or in a cluster with another sonorant consonant such as the /y/ in "banyan" and the /r/ in "snarl." Note that all of the semivowels can be the second member of an intervocalic sonorant consonant cluster since all of them can be prevocalic. However, as stated earlier, only the semivowels /l/ and /r/ can be postvocalic. Thus, of the semivowels, only /l/ and /r/ can be the first member of an intervocalic sonorant consonant cluster.

The detection strategy begins by finding all regions within an utterance which are voiced and sonorant. Next, as stated earlier, anchor points are placed within the voiced sonorant regions on the basis of significant energy change and significant formant movement. That is, dip detection is performed within the time functions representing the mid-frequency energies to locate all nonsyllabic sounds. Dip detection and peak-detection are performed on the tracks of F2 and F3 to extract some of the formant based properties possessed by one or more of the semivowels. The F2 dip detection algorithm marks sounds which are more "back" than adjacent segments. Thus, as the data of Section 3.2.2 show, the detection of this type of formant movement should find most /w/'s as well as many /l/'s and /r/'s. The F2 peak detection algorithm marks sounds which are more "front" than adjacent sounds. Thus, this algorithm should locate most of the /y/ glides. Most of the retroflexed /r/ and some labial /w/ sounds should be found from dip detection of F3. Finally, the F3 peak detection algorithm should locate many of the nonlabial and nonretroflexed semivowels /l/ aud /y/ since they usually have an F3 frequency greater than or equal to that of adjacent sounds. In addition, as the data of Section 3.2.2 show, /w/'s which are in a retroflexed environment mãy be detected in this way.

The results of the acoustic study of Chapter 3 are embedded in the different detection algorithms in other ways as well. Before marking a maximum or minimum in the energy and formant parameters, the amount of change is taken into consideration. In addition, for the formant dips and peaks, the frequency at which they occur must fall within an expected range of values. Thus, not all maxima and minima within these parameters are marked by the algorithms.

While the principle is the same, different detection algorithms were developed to find the sonorant-initial, sonorant-final and intersonorant semivowels. The results of some algorithms are used in other algorithms such that the detection of the semivowels follows a hierarchy. Because they can be detected most reliably, the intersonorant semivowels are detected first. The resulting anchor points are then used to detect the sonorant-final semivowels. Finally, the results from both the intersonorant and sonorant-final detection schemes are used to detect the sonorant-initial semivowels. Discussion of these different algorithms will follow this hierarchy.

Intersonorant Semivowels

A recursive dip detection algorithm (Mermelstein, 1975) was implemented to find minima in the mid-frequency energies and in the tracks of F2 and F3. Peak detection within the F2 and F3 waveforms is also performed by the dip detection algorithm by inverting the formant tracks. This algorithm marks minima which are surrounded by maxima. An example is shown in Figure 4.4. Since the intersonorant semivowels usually occur between vowels so that there are either V-C-V transitions or V-C-C-V transitions, one or more of the parameters will have this type of waveform shape with point *B* occurring within the semivowel, and points *A* and *C* occurring within the adjacent vowels. As indicated in Figure 4.4, the strength or depth of the dip is also computed. This value, which is labeled *d*, is the difference between the parameter value at point *B* and the smaller of the parameter values at the surrounding local maxima occurring at points *A* and *C*. The strength of the dips is used later in the integration of the dips for classification (see Section 4.3.2).

Some results obtained by using this algorithm are shown in Figure 4.5 which con-

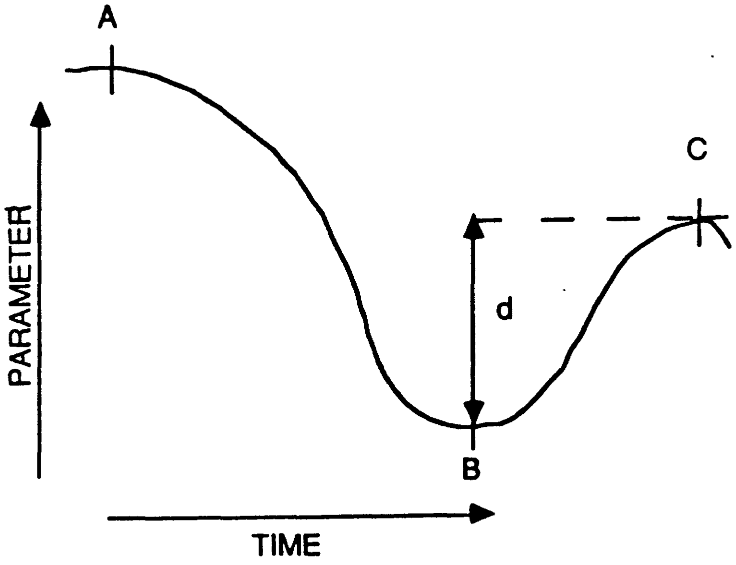


Figure 4.4: Illustration of intersonorant dip detection algorithm.

tains several displays relating to the word "willowy." The detected voiced sonorant region can be inferred from part a, which contains formant tracks that are computed only within this region. As can be seen from part c, the times of both of the F2 minima occurring within the intervocalic /l/ and /w/ segments are marked. The strengths of these dips are represented by the height of the spikes. Thus, while both semivowels have a dip in F2, the depth of the dip occurring within the /w/ segment is stronger than the depth of the dip occurring within the /l/ segment.

Although most sonorant-initial and sonorant-final semivowels are not detected by this algorithm, some acoustic events within these sounds may be marked if there is considerable movement in a parameter due to an adjacent nonsonorant consonant. An example of this phenomenon is shown in Figure 4.6 where the result of the F2 dip detection algorithm is shown for the word "dwell." In this case, due to the formant transitions between the /d/ and /w/, the prevocalic /w/ was detected by an intersonorant F2 dip.

On the basis of the anchor points placed by the energy dip detection algorithm, the locations of vowels are easily computed. Syllabic nuclei are determined by computing the time of maximum energy between the series of acoustic events including the beginning of the voiced sonorant region, the sequence of energy dips and the end of the voiced sonorant region. Both of these types of events within "willowy" are shown in parts d and e of Figure 4.5, respectively. Since both energy dips occurring within the intervocalic semivowels /l/ and /w/ are detected, the energy peaks occurring within

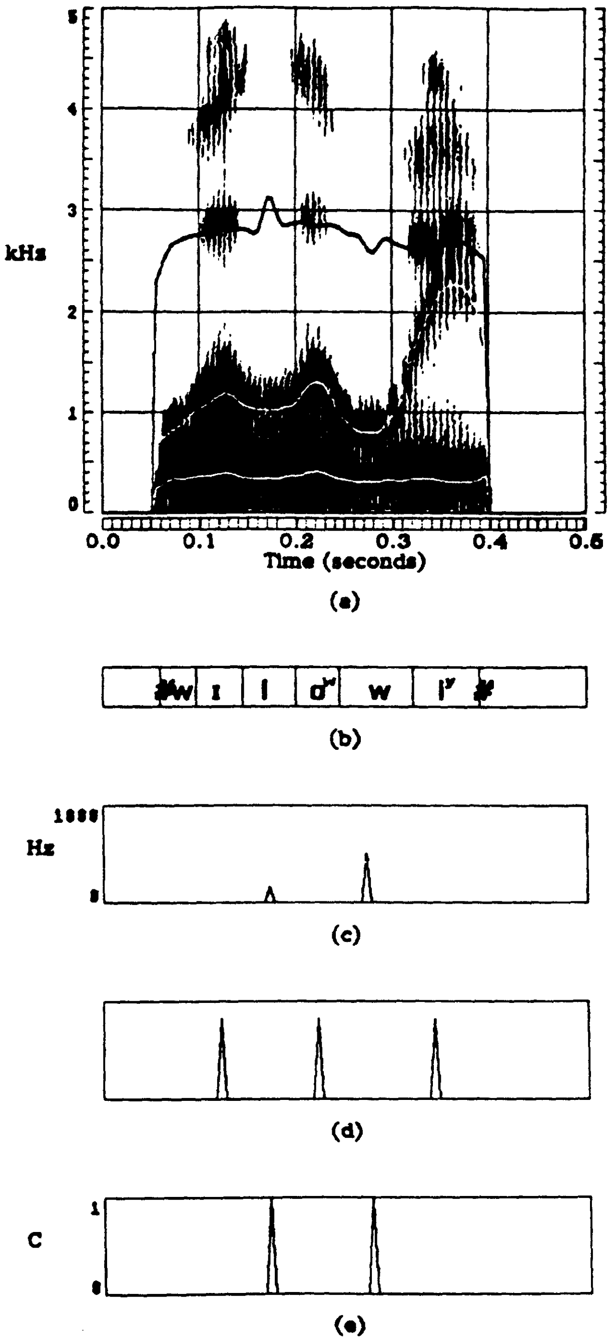


Figure 4.5: Results of Intersonorant dip detection in "willowy." (a) Wide-band spectrogram with formant tracks overlaid. (b) Phonetic transcription. (c) Location and depth of F2 dips. (d) Location of energy peaks (e) Location and confidence of energy dips.

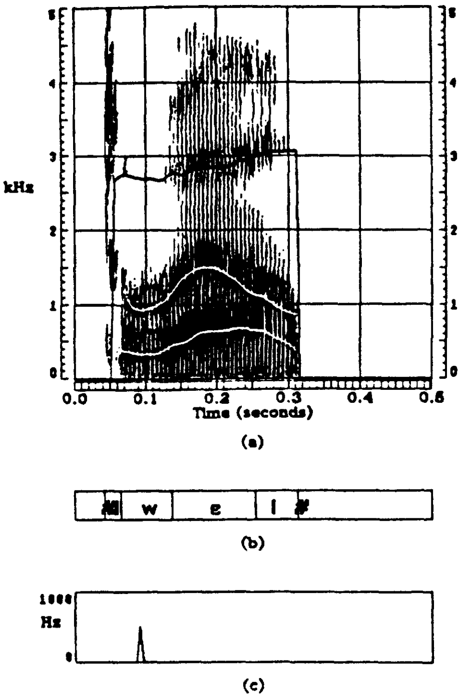


Figure 4.6: Result of Intersonorant F2 dip detection in "dwell." (a) Wide-band spectrogram with formant tracks overlaid. (b) Phonetic transcription. (c) Location and depth of F2 dip.

the vowels are also located. As is discussed below, the time of the energy maxima are used in both the sonorant-initial and sonorant-final semivowel detection algorithms.

Sonorant-Final Liquids

Of the semivowels, only the liquids /l/ and /r/ occur in postvocalic and, therefore sonorant-final positions. Thus, the F2 peak detection algorithm used to locate the /y/ glide is not used in this detection scheme.

The data of Section 3.2.2 show the type of formant movement indicative of a sonorant-final /l/ and /r/. If an /l/ is at the end of a voiced sonorant region, there is usually significant downward movement in F2 and/or significant upward movement in F3 from the preceding vowel into the /l/. In the case of a sonorant-final /r/, there is usually significant downward movement in F3 from the preceding vowel and possibly downward movement in F2 if the vowel is "front." As in the previous section, sonorant-final peak detection of F3 is performed by inverting the track of F3 and doing dip detection. Thus, the detection algorithm marks minima in waveforms whose shape at the end of voiced sonorant regions resembles the one shown in Figure 4.7. Points

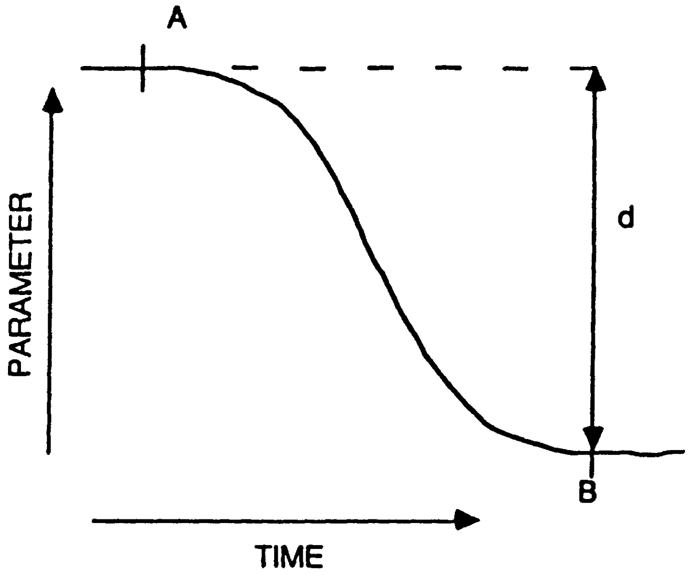


Figure 4.7: Illustration of sonorant-final dip detection algorithm.

A and *B* correspond to the times of the maximum and minimum formant or energy values within the vowel and following semivowel segments, respectively. The strength of the dip labeled *d* is simply the difference between the values occurring at these times.

To determine points *A* and *B* in a parameter, we need to monitor the movement of the waveform throughout the vowel and semivowel regions. Recall that energy maxima are computed once all the intersonorant energy dips are computed. Thus, the time of the last energy maximum within the voiced sonorant region corresponds to point *A* when the waveform is one of the mid-frequency energies. To determine point *A* in the formant tracks, we estimate the beginning and end of the vowel within which the last energy maximum occurs and compute the maximum formant value occurring between these times. The onset and offset waveforms are used for this purpose. More specifically, the vowel onset is taken to be the time at which there is the greatest rate of change in energy between the sound preceding the vowel and the energy peak occurring within the vowel. If an intersonorant energy dip indicating an intersonorant sonorant consonant precedes the energy peak within the vowel, then the beginning of the vowel is taken to be the time of the onset occurring between these events. However, if no intersonorant energy dip precedes the energy peak, then the beginning of the vowel is taken to be the time of the onset occurring between the beginning of the detected voiced sonorant region and the time of the energy peak. In cases where the vowel occurring before the sonorant-final liquid is preceded by a sonorant-initial semivowel or nasal which has not as yet been detected (recall that sonorant-initial dip detection is performed after sonorant-final dip detection), the vowel onset may be incorrectly estimated since the onset of the sonorant-initial consonant may be greater than the onset of the vowel. However, we have not found this to be a problem in the determination of point *A* and, therefore, in the detection of the sonorant-final liquids.

Similar to the vowel onset, the vowel offset is taken to be the time of the greatest rate of change in the offset waveform between the last energy maximum and the time occurring 10 msec before the end of the voiced sonorant region. The time of the vowel offset is also used to determine point *B* which is the time between this event and 10 msec before the end of the voiced sonorant region at which the minimum formant or energy value occurs.

Results obtained with this algorithm are shown in Figure 4.8 which contains several displays pertaining to the word "yell." As can be seen in parts d and e, estimates of the vowel onset and offset, which occur at 154 msec and 256 msec, respectively, appear to be reasonable. Thus, the movement of F2 and F3 between the /ɛ/ and following /l/ is detected. Both an F2 minimum and F3 maximum shown in parts f and g, respectively, are found within the /l/.

Sonorant-Initial Semivowels

The strategy used to detect sonorant-initial semivowels is based on a comparison between the beginning of a voiced sonorant region and the first vowel region. From the data presented in Chapter 3, we have made several observations. First, many word-initial semivowels have significantly less energy than the following vowel. Second, between a prevocalic /w/, /l/ or /r/ and the following vowel, F2 usually rises significantly. Third, between a prevocalic /r/ and the following vowel, F3 usually rises significantly from a value normally below 2000 Hz. Finally, following a prevocalic /y/, F2 and F3 fall gradually from a fronted position.

As before, peak detection in F2 and F3 is done by inverting the tracks and doing dip detection. Thus, if a semivowel is present, we expect one or more of the energy and formant parameters to have a waveform shape at the beginning of the detected voiced sonorant region which is similar to that shown in Figure 4.9. Point *A* is the time of the maximum parameter value within the first vowel in the voiced sonorant region. When the parameter is one of the bandlimited energies, this point will correspond to the first energy peak placed by the vowel detection program discussed above. As in the case

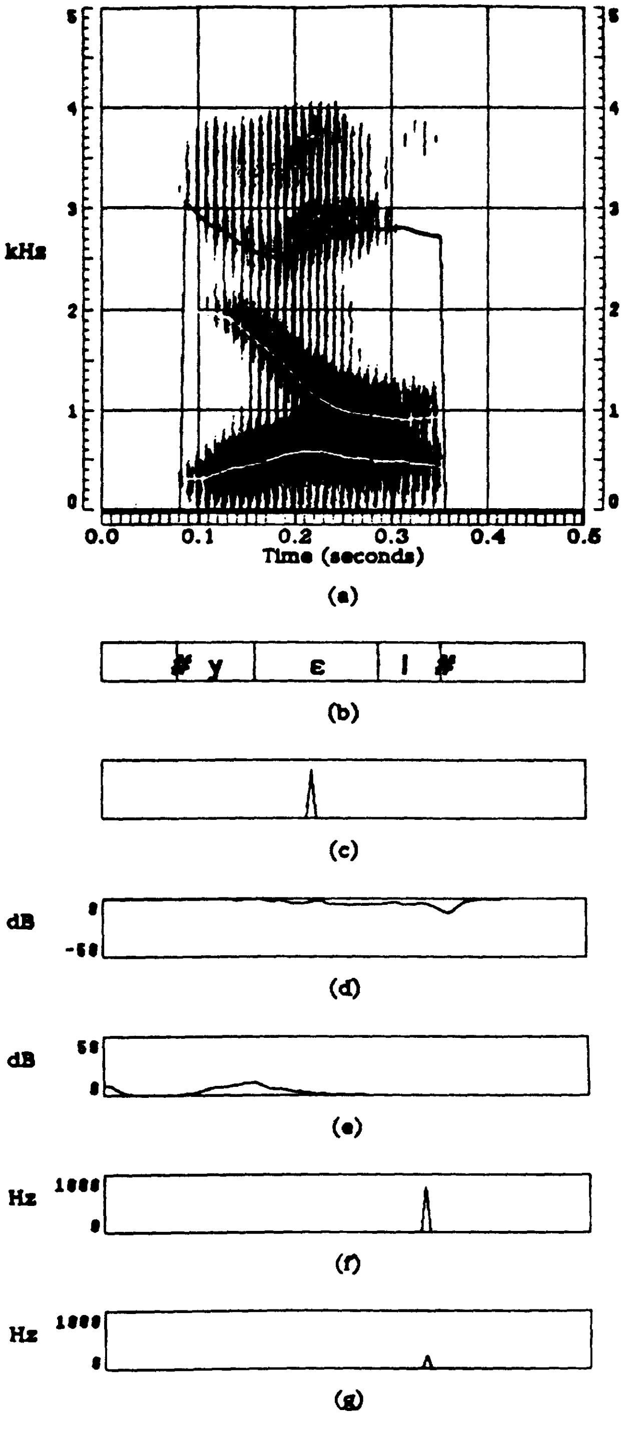


Figure 4.8: Results of sonorant-final dip detection in "yell." (a) Wide-band spectrogram with formant tracks overlaid. (b) Phonetic transcription. (c) Location of energy peak. (d) Offset waveform. The time of the vowel offset is estimated to be 256 msec. (e) Onset waveform. The time of the vowel onset is estimated to be 154 msec. (f) Location and depth of F2 dip. (g) Location and depth of F3 peak.

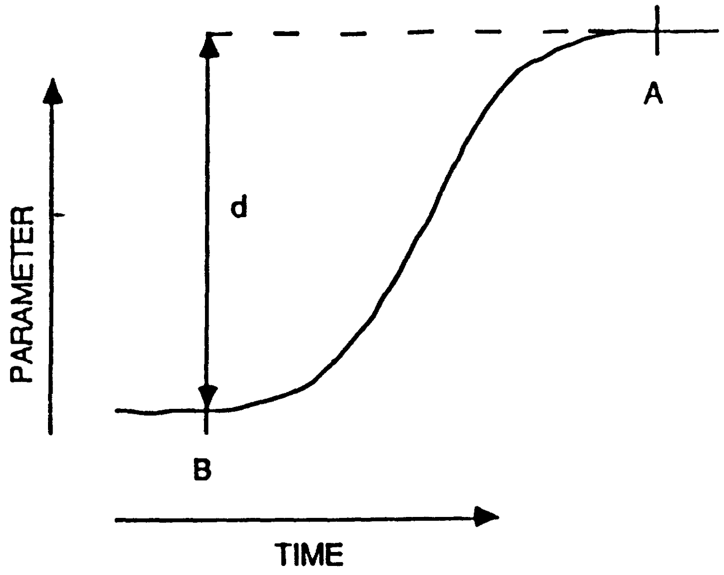


Figure 4.9: Illustration of the sonorant-initial dip detection algorithm.

of the detection of sonorant-final semivowels, when a formant track is the parameter, its movement throughout the first vowel must be monitored to determine point A. Again, the offset waveform is used to determine the end of the first vowel region which is taken to be the time of the offset occurring between the first energy peak and the following boundary. This boundary may be either an intersonorant energy dip, an acoustic event marked by one of the sonorant-final dip detection algorithms or, if none of these exists, the end of the voiced sonorant region.

Point B is the time 10 msec into the voiced sonorant region. Thus, if the difference d between the parameter values at points A and B is significant, point B is marked by a spike with a height of d.

Results obtained with this algorithm are shown in Figure 4.10 where several displays relating to the word "yell" are presented. As can be seen in parts d, e and f, F2 and F3 maxima and an energy minimum are marked in the sonorant-initial /y/.

4.3.2 Classification

Based on the type of acoustic events marked within the region of the detected sound(s), the classification step does two things. First, it extracts all of the acoustic properties from a region surrounding an anchor point selected from amongst the acoustic events. This process involves the computation of average F1, F2 and F3 frequencies which are based on the formant values at the time of the anchor point and

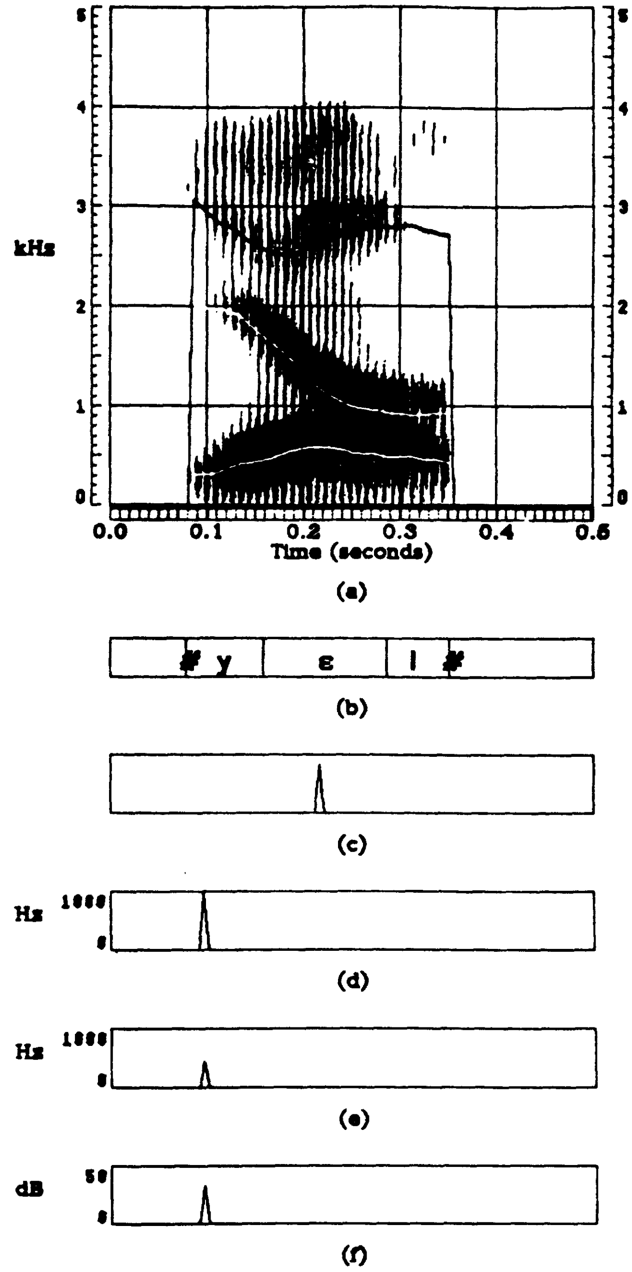


Figure 4.10: Results of sonorant-initial dip detection in "yell." (a) Wide-band spectrogram with formant tracks overlaid. (b) Phonetic transcription. (c) Location of energy peak. (d) Location and depth of F2 peak. (e) Location and depth of F3 peak. (f) Location and depth of energy dip.

the values occurring in the previous and following frames. In addition, F0 is computed by averaging together "reasonable" estimates occurring throughout the utterance (as mentioned in Chapter 3). From these values, the formant-based properties listed in Table 4.2 are computed and quantified. The anchor point is also used to extract the acoustic correlate of the feature *stop* which characterizes the rate of spectral change between the detected sound and surrounding segments. In particular, if the anchor point is preceded by an energy maximum (which should occur within the preceding vowel), the offset between these events is extracted and quantified. Similarly, if the anchor point is followed by an energy maximum (which should occur in the following vowel), the onset between these events is extracted and quantified. With the quantified properties determined, the second step in this recognition process decides which semivowel rules should be invoked.

The implementation of these steps differs somewhat depending upon whether a detected sound is thought to be sonorant-initial, intersonorant or sonorant-final. Thus, we discuss separately below the classification strategies for these contexts. Finally, we end this section with a discussion of the semivowel rules.

Sonorant-Initial Classification Strategy

A flow chart of the strategy used to classify sounds detected by one or more sonorant-initial dips is shown in Figure 4.11. Basically, the algorithm starts by trying to determine what, if any, acoustic events have been marked between the beginning of the detected sonorant region and the first energy peak which should occur within the first vowel. As implied in the flow chart, the determination of what acoustic events have been marked follows a hierarchy. This is so because some events, more so than others, narrow the choice(s) of semivowels. For example, if an F2 peak is marked, then, of the semivowels, we will only investigate the possibility of the sound being a /y/. Thus, branch 3 is implemented so that the F2 peak is used as the anchor point and only the /y/ rule is applied.

On the other hand, if an F2 dip is marked, then the detected sound could be a /w/, /l/ or /r/. In this case, branch 1 is implemented. To further narrow the choices, the algorithm looks to see what other events, if any, have been detected. For example, if in addition to an F2 dip an F3 peak is marked, then the F2 dip is the anchor point and the /r/ rule is not invoked. Instead, the /l/ rule is applied and, as indicated by the bidirectional arrow, the /w/ and /w-l/ rules may also be applied. Recall that the

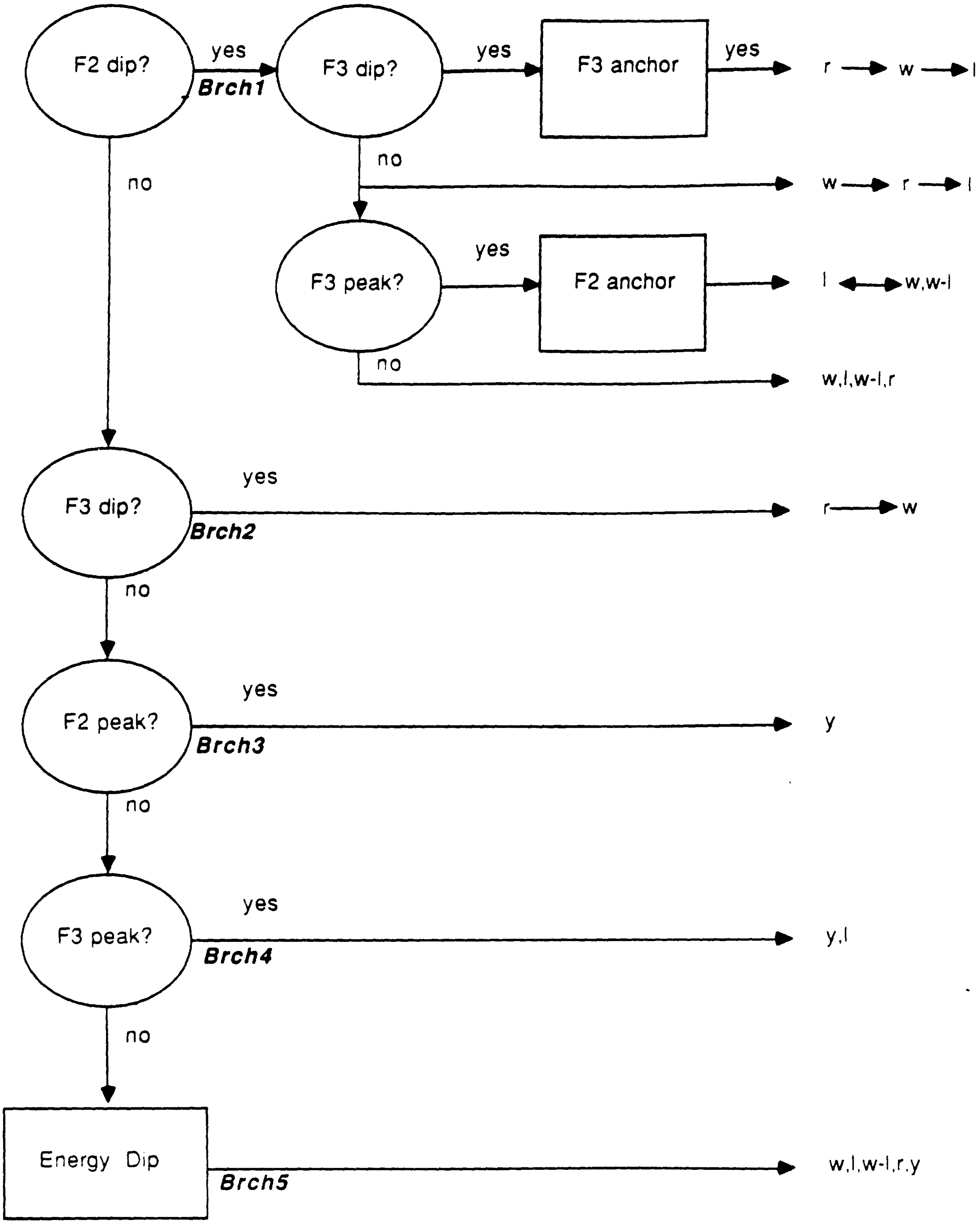


Figure 4.11: Flow chart of the sonorant-initial classification strategy.

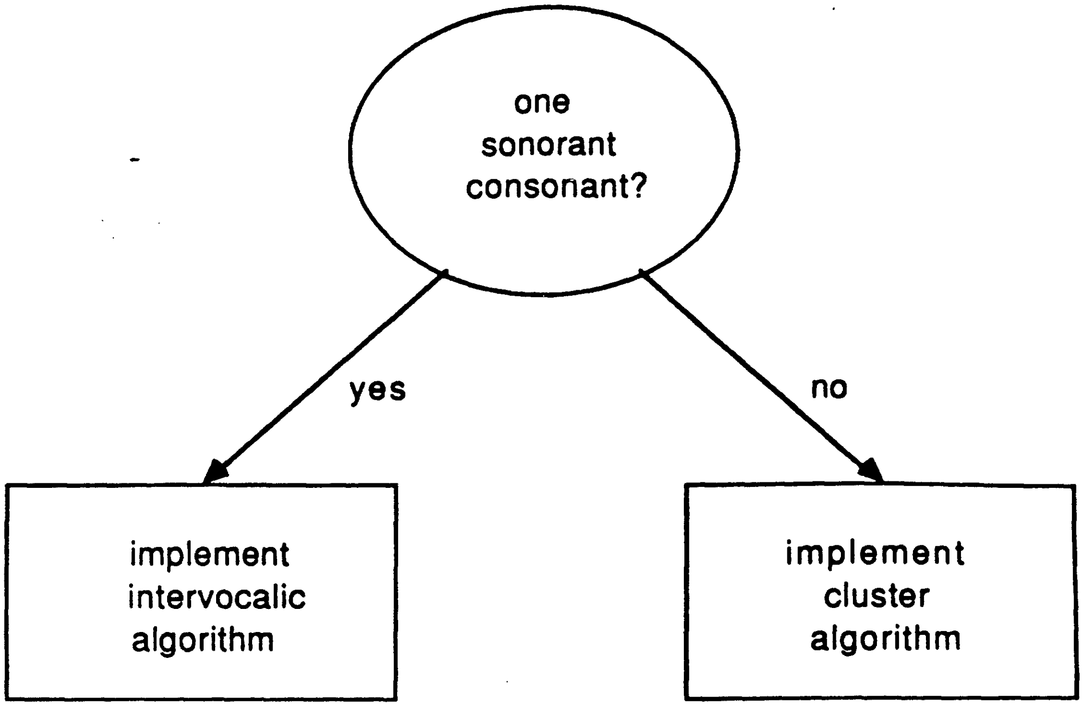


Figure 4.12: Flow chart of the intersonorant classification strategy.

data of Section 3.2.2. show that /w/'s which are adjacent to retroflexed sounds have a higher F3 frequency. Thus, the latter rules are applied if it is determined, from an analysis of a small region surrounding the first energy peak, that the first vowel is retroflexed.

Note that several unidirectional arrows also occur in the flow chart such as the ones in the path of branch 1 which contains an F2 dip and F3 dip. The first arrow implies that the /r/ rule is applied first and that the /w/ rule is invoked only if the sound is not classified as an /r/. Similarly, the second arrow states that if the sound is not classified as an /r/ or a /w/, then the /l/ rule is applied.

Intersonorant Classification Strategy

This classification strategy is more complicated than the others since an intersonorant dip region (loosely defined by the energy maxima surrounding the energy dip) may consist of one or two sonorant consonants. Thus, as Figure 4.12 shows, the first step in this process is the determination of whether one or two sonorant consonants are in the dip region.

The answer to this question is based on two types of acoustic cues. The first type of cue has to do with duration. Recall that the data of Section 3.2.6 show that the difference in time between the offset and onset surrounding an intersonorant energy dip is usually much longer when two sonorant consonants are present and the first consonant is a nasal than when either one sonorant consonant is present or two sonorant consonants are present and the first one is either /r/ or /l/ (recall that /w/ and /y/ cannot be the first member of an intersonorant cluster). Thus, to differentiate between the latter events and, therefore, determine if there is either one sonorant consonant or a liquid followed by another sonorant consonant, the algorithm looks to see if an F3 dip (indicating an /r/) or either an F3 peak or F2 dip (indicating an /l/) occurs between the surrounding energy maxima, but either before or just after the offset. Examples of these pattern of events within the words "harmonize" and "stalwart" are shown in Figure 4.13. As can be seen, only the /m/ in the /rm/ cluster of "harmonize" occurs between the offset and onset at 274 msec and 334 msec, respectively. The presence of the /r/ is indicated by the strong F3 dip shown in part f which occurs just before the offset. Similarly, the presence of the /l/ in the /lw/ cluster in "stalwart" is indicated by an F3 peak occurring just before the offset at 352 msec.

If the algorithm determines that only one consonant is present in the dip region, then the intervocalic classification strategy shown in Figure 4.14 is implemented. This strategy is very similar to the one which classifies sounds detected by sonorant-initial acoustic events. In this case, however, the determination of acoustic events does not follow a hierarchy, except that the algorithm favors formant dips/peaks over energy dips. Instead, the algorithm determines the strongest acoustic event (recall that the strength of a dip is determined in addition to the time at which it occurs). Thus, if an F2 dip and F3 dip both occur, but the F3 dip is stronger, then branch 2 in Figure 4.14 is implemented.

If the algorithm determines that two sonorant consonants are present in the dip region, then the cluster classification strategy shown in Figure 4.15 is implemented. The first step in this process is the determination of whether the first sonorant consonant in the cluster may be a postvocalic liquid. If so, path 1 is followed. As can be seen, after either the /r/ or /l/ rule is applied to the sound detected by the formant dip/peak, the intervocalic classification strategy discussed above is used to classify the sound occurring between the offset and onset.

Path 2 is followed when the algorithm suspects that two sonorant consonants occur between the offset and onset. If this is so, then we know that the first consonant is

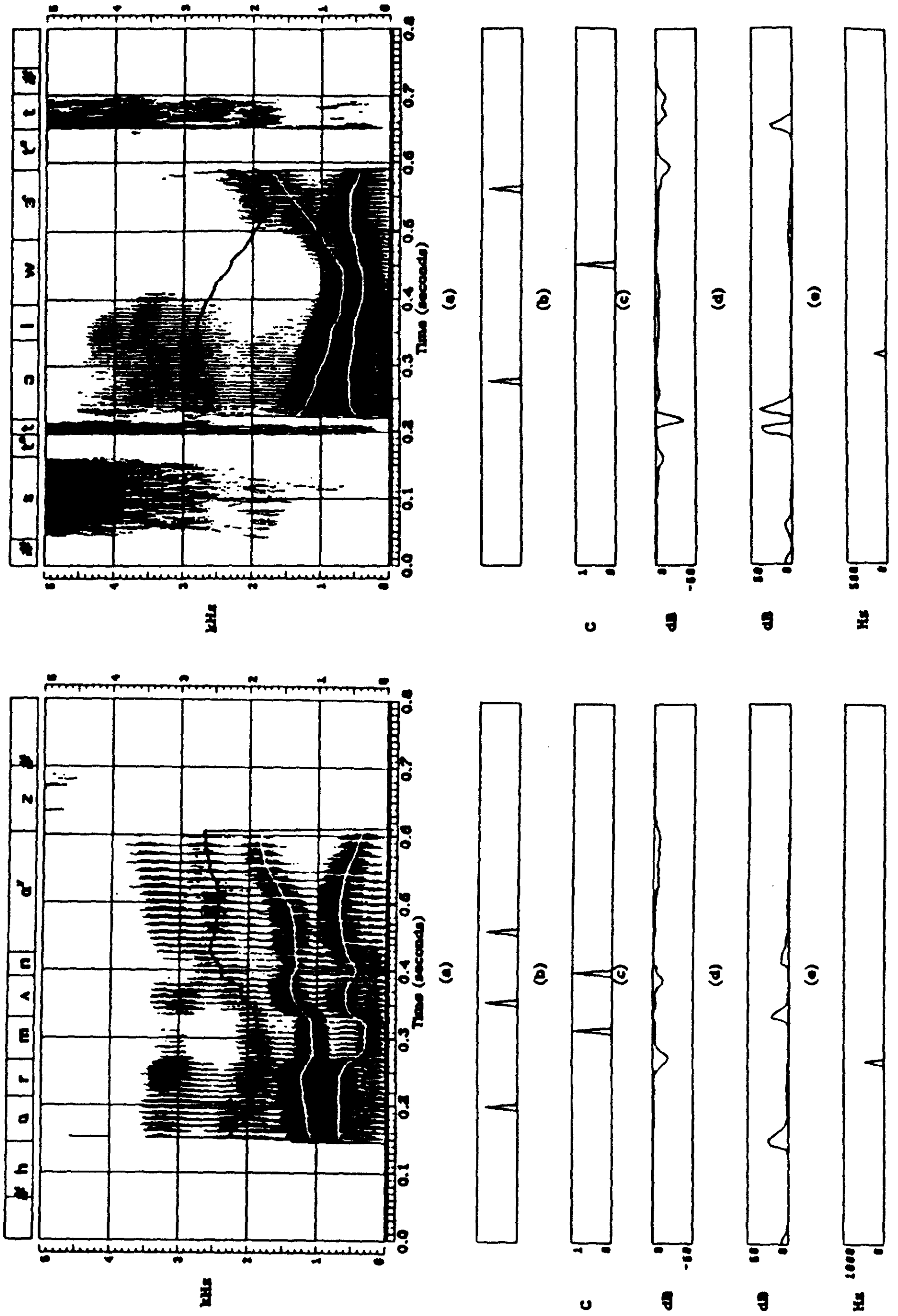


Figure 4.13: Pattern of events expected when /r/ or /l/ are postvocalic and in an intersonorant cluster. (a) Wide band spectrograms of the words "harmonize" and "stalwart" with formant tracks overlaid and phonetic transcriptions on top. (b) Location of energy peaks. (c) Location and confidence of energy dips. (d) Onset waveform. (e) Offset waveform. (f) Location and depth of F3 dip in "harmonize" and F3 peak in "stalwart."

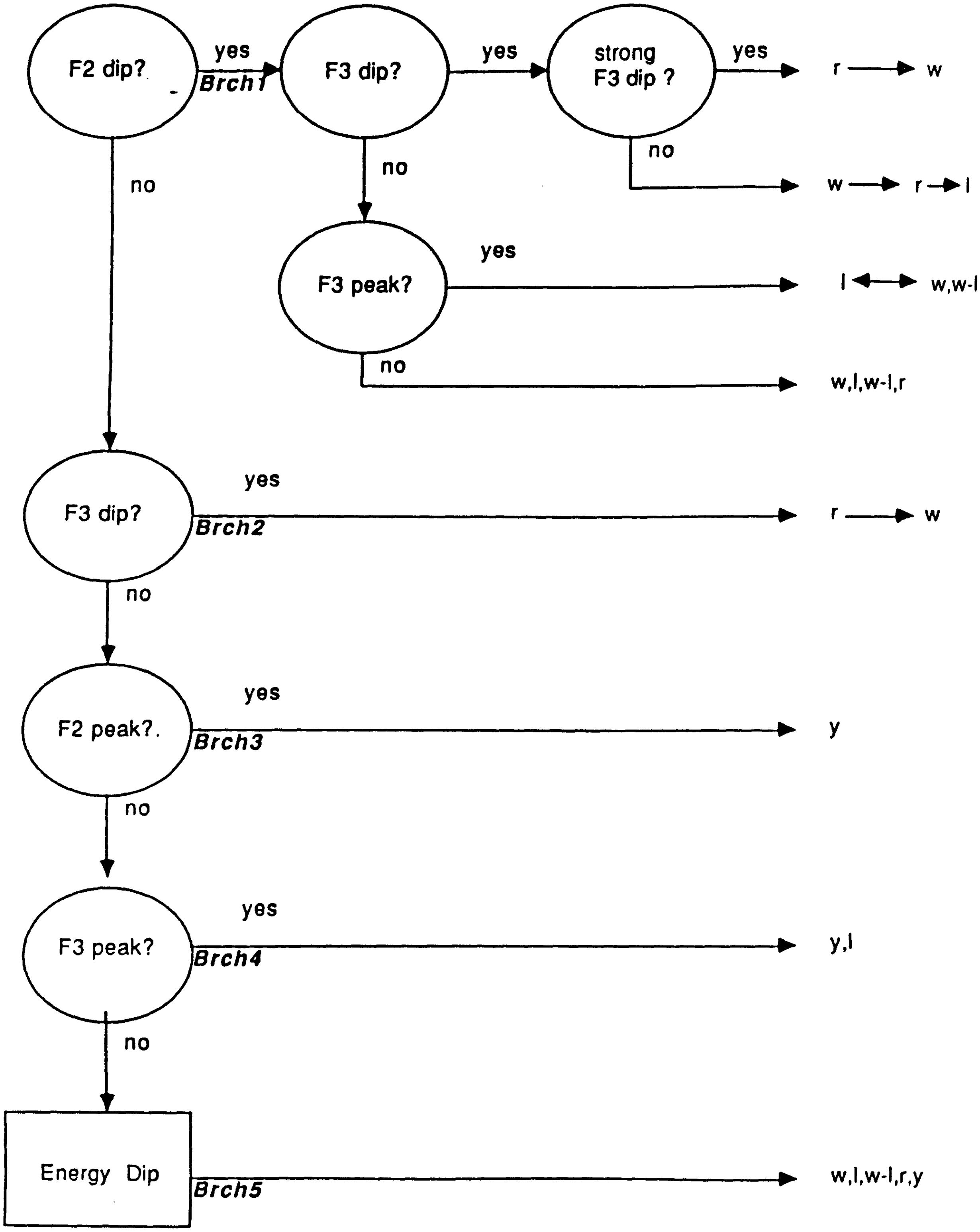


Figure 4.14: Flow chart of the intervocalic classification strategy.

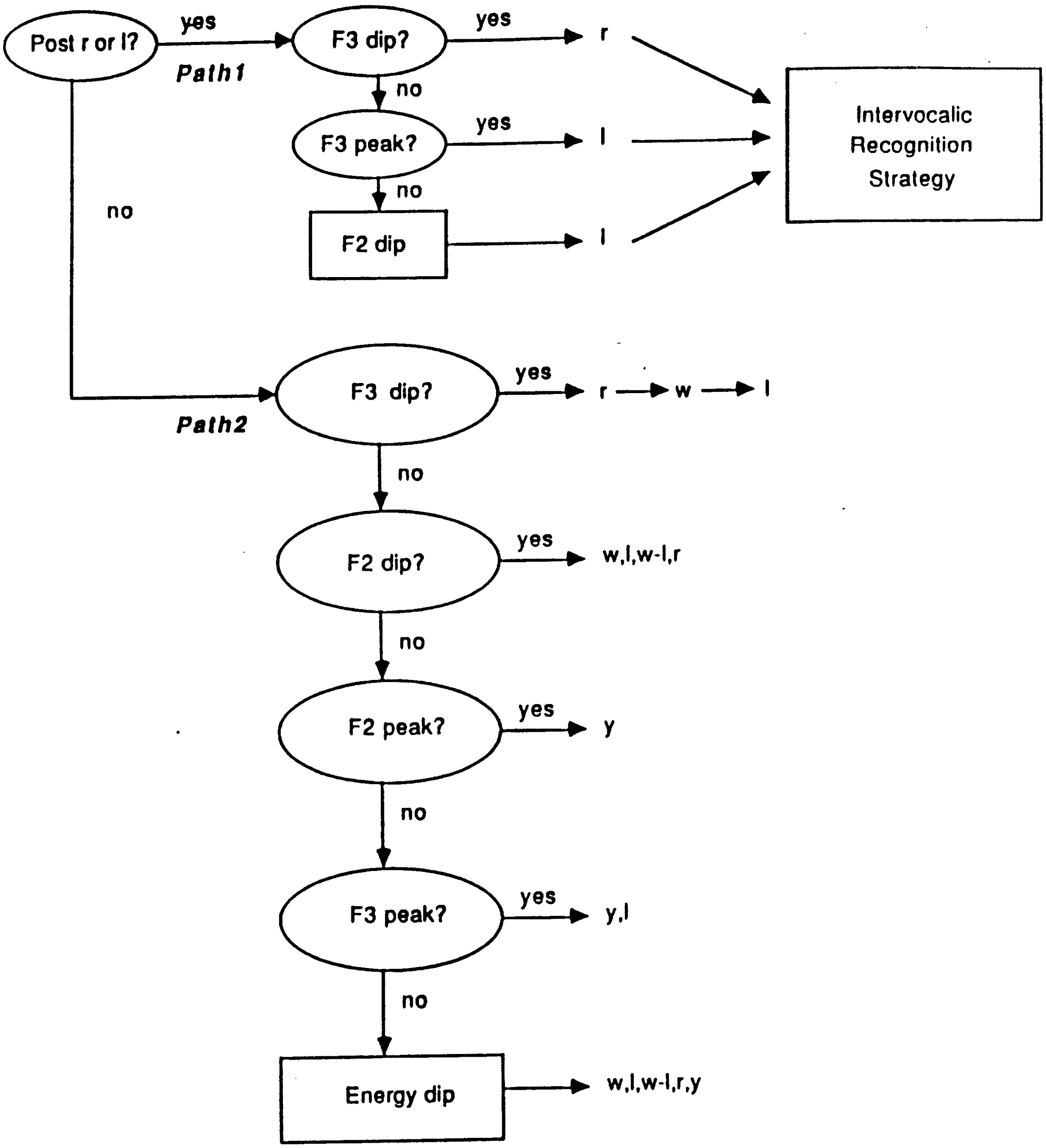


Figure 4.15: Flow chart of the cluster classification strategy.

not a semivowel and that the second consonant is either a nasal or semivowel. Thus, we only want to classify the second consonant. To try to guarantee that only the second consonant in the cluster is classified, the algorithm selects the last acoustic event occurring in the energy dip region. This is the question which is being asked in Path 2. For example, if the last acoustic event is an F3 peak, then the /y/ and /l/ rules are applied.

Sonorant-Final Classification Strategy

The classification strategy for acoustic events occurring in a sonorant-final region (loosely defined as the interval between the last energy maximum and the end of the voiced sonorant region) is shown in Figure 4.16. As can be seen, this process is straightforward since, of the semivowels, only /l/ and /r/ can occur in a sonorant-final position. The hierarchy implied is not crucial except that branches 1 and 2, because a dip or peak in F3 distinguishes between the liquids, should be implemented before the lower ones.

Rules

While the thresholds used to quantify the extracted properties are always the same, the rules which are applied to integrate them for identification of the semivowels are dependent upon context. The rules for the different contexts are compared in Tables 4.5, 4.6 and 4.7. As stated above, there is a /w-l/ rule for a class which is either /w/ or /l/. This category was created since, as the acoustic study discussed in Chapter 3 shows, /w/ and /l/ are acoustically very similar.

In the fuzzy logic framework, addition is analogous to a logical "or" and the result of this operation is the maximum value of the properties being considered. Multiplication of two or more properties is analogous to a logical "and." In this case, the result is the minimum value of the properties being operated on. Since the value of any property is between 0 and 1, the result of any rule must also be between 0 and 1. We have chosen 0.5 to be the dividing point for classification. That is, if the sound to which a semivowel rule is applied receives a score greater than or equal to 0.5, it will be classified as that semivowel.

Although the rules are similar across contexts, well known acoustic differences due to allophonic variations are captured in the rules. For example, compare the prevocalic and postvocalic /l/ rules. The rule for a prevocalic /l/ allows for the possibility of an

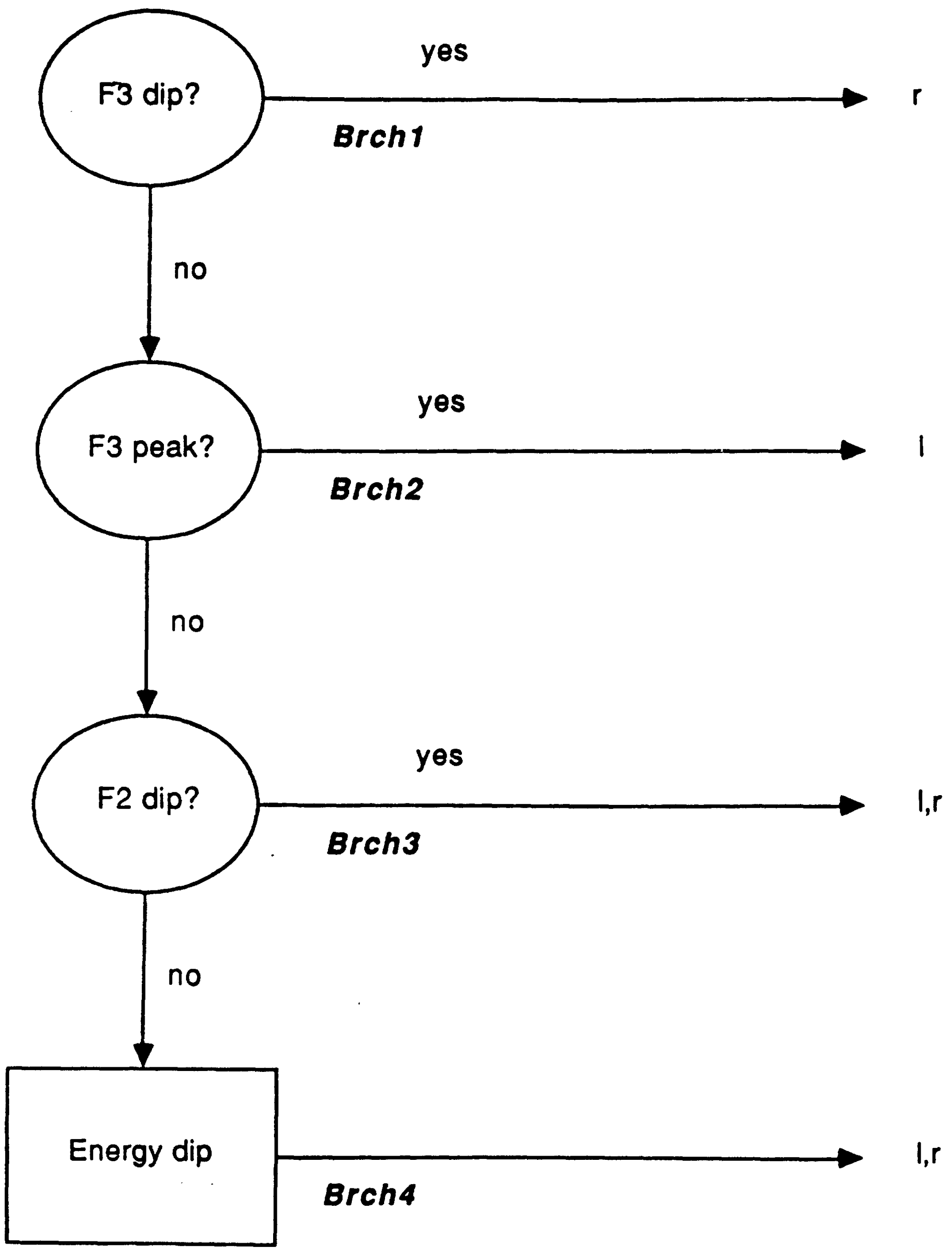


Figure 4.16: Flow chart of the sonorant-final classification strategy.

Table 4.5: Prevocalic Semivowel Rules

/w/	=	(very back) + (back)(high + maybe high)(gradual onset) (maybe close F2 F3 + not close F2 F3)
/l/	=	(back + mid)(gradual onset + abrupt onset)(maybe high + nonhigh + low) (maybe retroflex + not retroflex) (maybe close F2 F3 + not close F2 F3)
/w-l/	=	(back) (maybehigh) (gradual onset)(maybe close F2 F3 + not close F2 F3)
/r/	=	(retroflex) (close F2 F3 + maybe close F2 F3) + (maybe retroflex) (close F2 F3) (gradual onset) (back + mid) (maybe high + nonhigh + low)
/y/	=	(front)(high + maybe high) (gradual onset + abrupt onset)

abrupt rate of spectral change between the detected sound and the following vowel. However, the rule for a postvocalic /l/ requires that the rate of spectral change between the detected sound and the preceding vowel be gradual. In addition, the closer spacing between F2 and F1 for a postvocalic /l/ as oppose to a prevocalic /l/ is also expressed. Whereas the rule for a postvocalic /l/ allows for the sound to be "very back," the rule for a prevocalic /l/ does not. Instead, to classify as an /l/, the detected sound must be either "back" or "mid."

Note that the fuzzy logic framework provides a straightforward mechanism for distinguishing between primary and secondary cues. For example, in the /w/ rules, the property "very back" is primary whereas the other cues are secondary. That is, if the sound has the property "very back," it will be classified as a /w/ regardless of the other properties. Otherwise, to be classified as a /w/ the sound needs to possess the properties "back," "gradual," and either "high" or "maybe high." Likewise, regardless of the value of any other properties, a sound which has the properties "retroflex" and a "close F2 and F3" or a "maybe close F2 and F3" (the postvocalic /r/ does not allow the last property), will be recognized as an /r/.

Table 4.6: Intersonorant Semivowel Rules

/w/	=	(very back) + (back)(high + maybe high)(gradual onset)(gradual offset) (maybe close F2 F3 + not close F2 F3)
/l/	=	(back + mid)(maybe high + nonhigh + low) (gradual onset + abrupt onset)(gradual offset + abrupt offset) (maybe retroflex + not retroflex) (maybe close F2 F3 + not close F2 F3)
/w-l/	=	(back) (maybehigh) (gradual onset) (gradual offset) (maybe close F2 F3 + not close F2 F3)
/r/	=	(retroflex) (close F2 F3 + maybe close F2 F3) + (maybe retroflex) (close F2 F3) (gradual onset) (gradual offset) (back + mid)(maybe high + nonhigh + low)
/y/	=	(front)(high + maybe high) (gradual onset) (gradual offset)

Table 4.7: Postvocalic Rules

/l/	=	(very back + back) (gradual offset) (not retroflex) (not close F2 F3)(maybe high + nonhigh + low)
/r/	=	(retroflex) (close F2 F3) + (maybe retroflex) (close F2 F3)(maybe high + nonhigh + low) (back + mid) (gradual offset)

4.3.3 Summary

In summary, we have divided the control strategy into two procedures: detection and classification. In the detection process, certain acoustic events (minima and maxima) which correspond to particular acoustic properties are automatically detected from selected parameters. In the classification process, these acoustic events are used in two ways. First, on the basis of their relative strengths and the time of their occurrence, they define a small region from which all of the acoustic properties for features are extracted. Second, once the properties are quantified, the acoustic events are used to decide which semivowel rule(s) will integrate the properties for classification of the detected sound.

Chapter 5

Recognition Results

5.1 Introduction

In this chapter, we evaluate the performance of the recognition procedures presented in Chapter 4. The detection and classification results are given separately for each of the data bases described in Chapter 2. The data base used to develop the recognition system is referred to as Database-1. Database-2 refers to the words contained in Database-1 which were spoken by new speakers. Finally, Database-3 refers to the sentences taken from the TIMIT corpus.

Recall that, whereas errors in the formant tracks of the words in Database-1 were corrected, those in the formant tracks of the utterances in Database-2 and Database-3 were not. Consequently, we have excluded from the recognition results those semivowels which were not tracked correctly and words which were not tracked at all (see the performance results for the formant tracker in Section 2.2.3).

In addition to overall recognition results for the data bases, separate results are given for the sonorant-initial, intersonorant and sonorant-final semivowels. To further establish the influence of context, additional divisions within these broad categories are sometimes made.

Before presenting the recognition data, we shall discuss several key issues that have a bearing on the understanding of them. These issues include the criteria used for tabulating the detection and classification results, the effects of phonetic variability due to such phenomena as stress and devoicing, and problems with some of the recognition parameters.

Finally, we will conclude this chapter with a comparison of the recognition system developed in the thesis and some earlier acoustic-phonetic front ends for which semivowel recognition results have been published. Unfortunately, we do not know of any statistically based recognition system for which recognition results for the semivowels have been published. Thus, we are not able to compare the performance of systems based on the different approaches.

5.2 Method of Tabulation of Results

A semivowel is considered detected if an energy dip and/or one or more formant dips and/or peaks is placed somewhere between the beginning (minus 10 msec) and end (plus 10 msec) of its hand transcribed region by any of the detection algorithms. The 10 msec margin, which was chosen arbitrarily, did not always include effects of the semivowels on what is considered to be the neighboring phoneme in the transcription. Thus, for about 1% of the semivowels, further corrections were made when tabulating the detection results. For example, consider the word "choleric" shown on the left side of Figure 5.1. Based on the above criterion, the F3 peak in part e occurs within the intervocalic /l/, but the first F2 dip in part d occurs in the preceding /ə/. However, it is clear that the fall of F2 from its maximum value within the /ə/ is due to the influence of the /l/. Thus, when tabulating the detection results, the F2 dip is considered to be in the /l/ and not in the /ə/.

In contrast with this example, consider the word "harlequin," shown on the right side of Figure 5.1. As can be seen in part d, an F3 dip is detected at the beginning of the sonorant region. Based on the stated criterion, the F3 dip does not occur within the /r/ segment. However, as in the previous example, this dip is also clearly due to the influence of the semivowel. Nevertheless, since it does not occur close to the hand transcribed /r/ region, but occurs at the beginning of the vowel, the dip is not assigned to the /r/. Thus, the results will state that the /r/ was not detected.

On the other hand, if the /ɑ/ in this example is recognized as an /r/, the recognition results will say that the /r/ was correctly classified and the /ɑ/ will not be included in the list of vowels misclassified as /r/. This disparity between the detection and classification results points to the problem in present transcription standards which do not allow for the overlapping of phonetic sounds. That is, we do not consider it an error if the sonorant-initial recognition strategy rather than the intersonorant recognition strategy classifies the /r/. As is the case in this example and as was discussed in

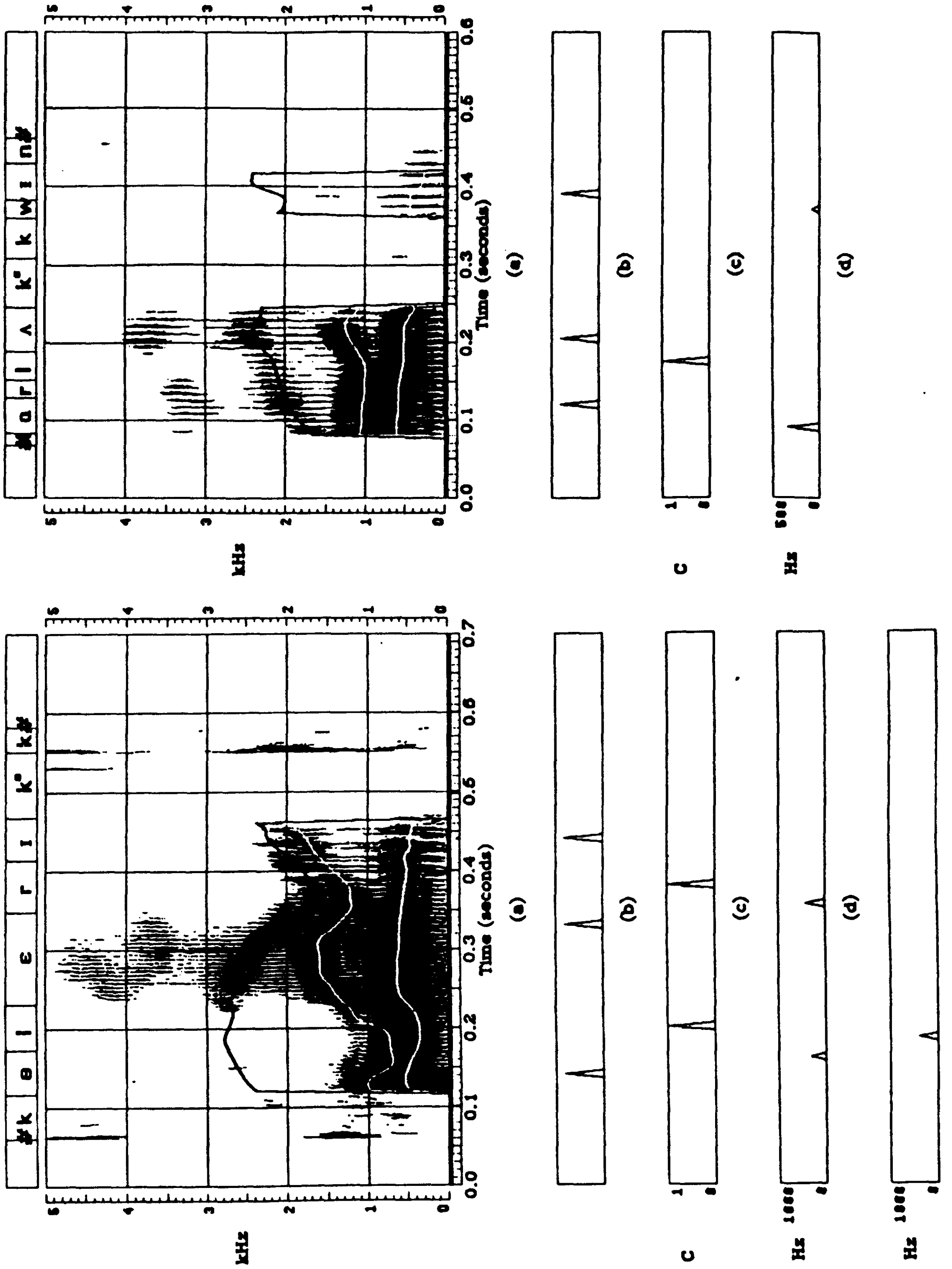


Figure 5.1: An illustration of acoustic events marked within "choleric" (on left) and "harlequin" (on right). (a) Wide band spectrogram with formant tracks overlaid. (b) Location of energy peaks. (c) Location and confidence of energy dips. (d) Location and depth of F2 dips in "choleric" and F3 dip in "harlequin." (e) Location and depth of F3 peak in "choleric."

Section 3.3, the features of an /r/ in this context may overlap completely with the preceding vowel. In this example, the underlying /ɑ/ and following /r/ segments are realized as an r-colored /ɑ/. Thus, in this sense, the /r/ is sonorant-initial. However, by allowing this "disorder" (or more appropriately "no order" since ideally this sound should be recognized as having the features of an /ɑ/ and an /r/) at the acoustic level, the unraveling of this r-colored segment into a vowel followed by an /r/ as opposed to a vowel preceded by an /r/ must occur at or somewhere before lexical access. Ideas concerning this mapping are discussed in chapter 6.

5.3 Effects of Phonetic Variability

The detection results are affected by phonetic variability due to stress and devoicing. Shown in Figure 5.2 are examples of unstressed semivowels. Formant tracks are given in the figure since some formants within the semivowels are not visible from the spectrogram. As can be seen, there appears to be little or no acoustic evidence for the /l/ in "luxurious" and the /y/ in "ukulele." Thus, neither of these semivowels is detected. This result is not surprising since perceptual findings (Cutler and Foss, 1977) have shown that acoustic cues of phonetic segments in unstressed syllables are not as salient as they are in stressed syllables. In fact, on the basis of this finding and their own work regarding lexical constraints imposed by stressed and unstressed syllables, Huttenlocher and Zue (1983) concluded that recognition systems may not need to be very concerned with the correct identification of phonetic segments in unstressed syllables.

In addition to some unstressed semivowels, devoiced and some partially devoiced semivowels are also undetected by the recognition system. Examples of such semivowels are shown in Figures 5.3 and 5.4. As can be seen, the /l/ in "clear," the /w/ in "swollen" and the first /r/ in "transcribe" are all considerably devoiced. As a result, they are not detected by the recognition system. Similarly, the /w/ in "mansuetude" and the prevocalic /l/ in "incredulously" are partially devoiced. In addition, these semivowels are unstressed. While there is enough formant movement so that the latter sounds are detected, the transitions are not sufficient for a correct classification. To recognize such semivowels, information in the preceding nonsonorant region is also needed. For example, the pencil-thin vertical line occurring above 5 kHz and between the /s/ and the following /l/ on the spectrogram of "incredulously" corresponds to

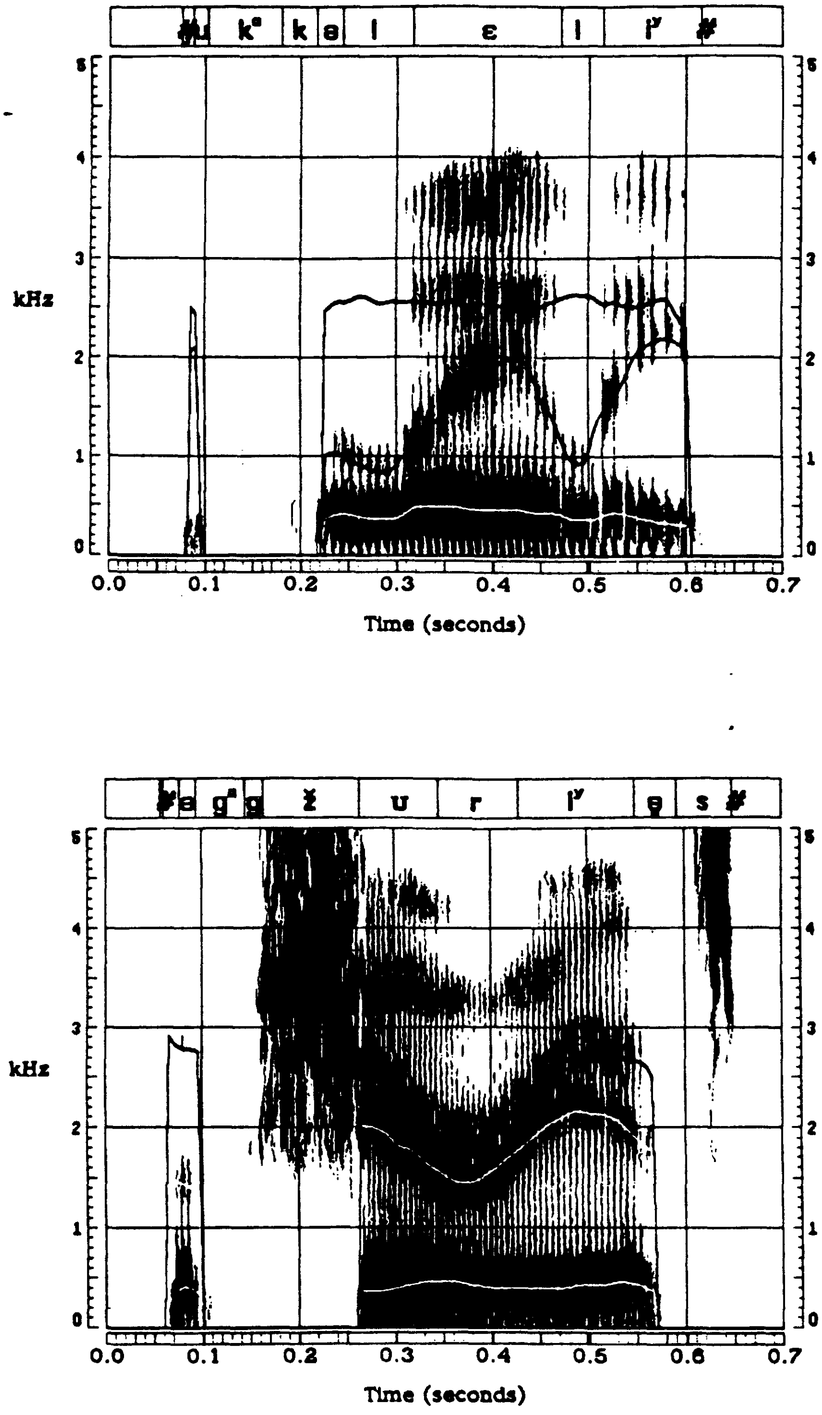


Figure 5.2: Wide band spectrogram with formant tracks overlaid of the words "ukulele" and "luxurious" which contain the unstressed, word-initial semivowels /y/ and /l/, respectively.

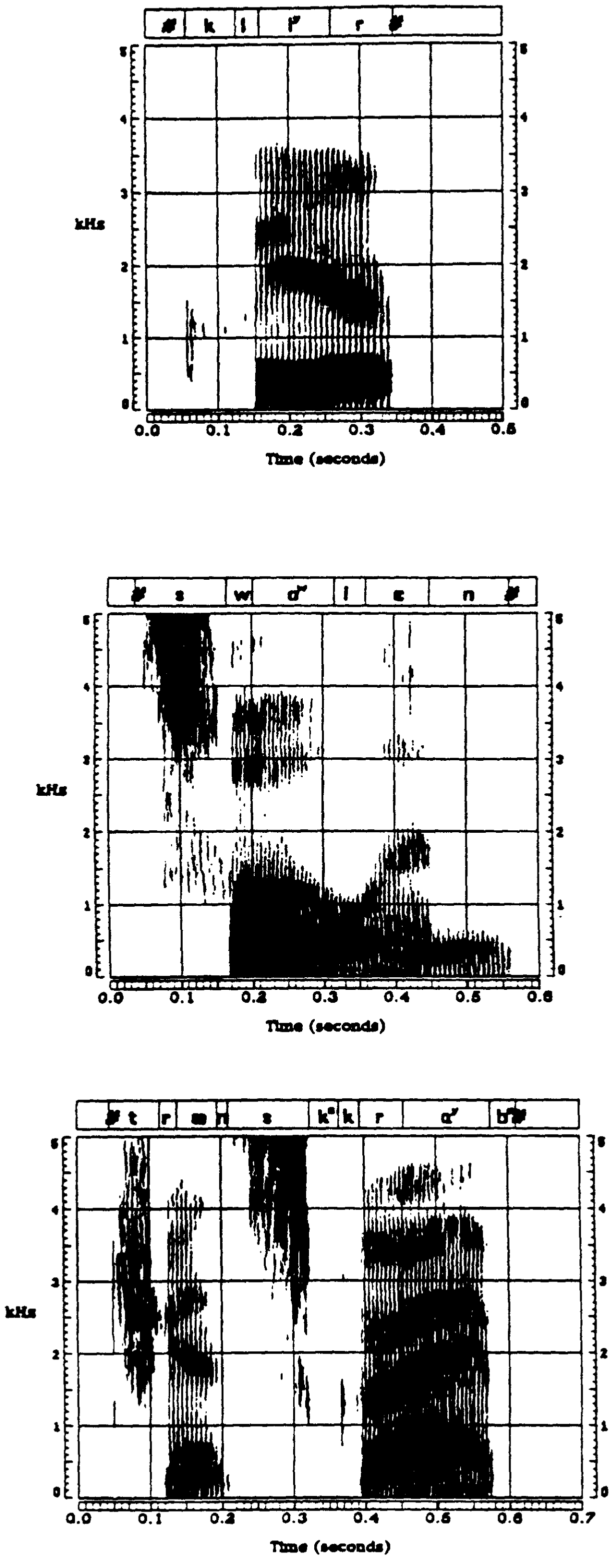


Figure 5.3: Wide band spectrograms of the words "clear," "swollen" and "transcribe" which contain a devoiced /l/, /w/ and /r/, respectively.

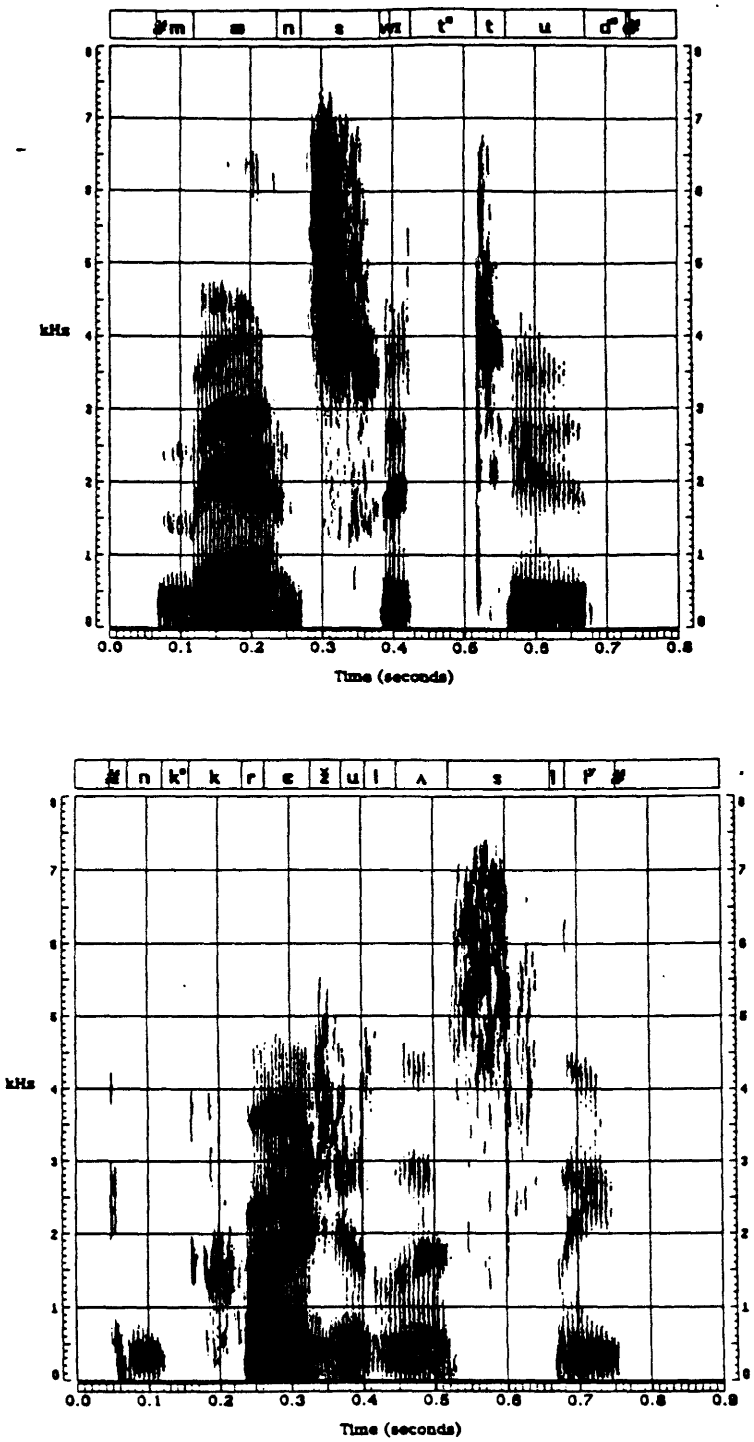


Figure 5.4: Wide band spectrogram of the words "mansuetude" and "incredulously." Both words contain unstressed, partially devoiced semivowels.

the lateral release of the /l/ (Zue, 1985). In addition, on the spectrogram of the word "mansuetude," the low frequency frication seen in the /s/ just below the starting point of F2 in the following voiced sonorant region is often referred to as a "labial tail" and is characteristic of a devoiced /w/. However, since analysis in the nonsonorant regions of an utterance is outside the scope of the thesis, semivowels such as these may not be detected. Recall that devoiced semivowels are not a part of our recognition task. However, since some words in the data bases contain semivowels which are in clusters with unvoiced consonants and since devoiced allophones and voiced allophones are transcribed with the same phonetic symbols, the detection and classification results for devoiced semivowels are included in the recognition data.

5.4 Parameter Evaluation

An evaluation of the voiced sonorant detector shows that, in a few instances, very weak sounds are excluded from the detected voiced and sonorant regions. Examples of this phenomenon are shown in Figures 5.5 and 5.6. As can be seen from the overlaid formant tracks which are extracted only in the detected voiced sonorant regions, the middle portion of the intervocalic /w/'s are excluded from the voiced sonorant regions. If we use the bandlimited energy from 200 Hz to 700 Hz, the difference (in dB) between the maximum energy within the utterance and the minimum energy within the /w/ is 37 dB for "bewail" and 41 dB for "bailiwick." As can be seen from the spectrograms, the /w/'s also have very little energy below 200 Hz. These results suggest that the /w/'s are produced with a constriction which is too narrow for them to be sonorant. Instead, they are produced as obstruents. Thus, their exclusion from the sonorant regions is reasonable.

Even though the intervocalic /w/'s shown in Figure 5.5 are partially excluded from the detected voiced sonorant region, they are still recognized. In each instance, enough of the /w/ is included in the following voiced sonorant regions so that it is detected and classified by the sonorant-initial recognition strategy.

While the exclusion of portions of the /w/'s in Figure 5.5 did not affect their recognition, the partial or complete exclusion of other semivowels from the detected voiced sonorant regions did cause them to be undetected and, therefore, unrecognized. Examples of such semivowels are shown in Figures 5.6 and 5.7. As can be seen in Figure 5.6, the last syllable in the word "harlequin," which contains a prevocalic

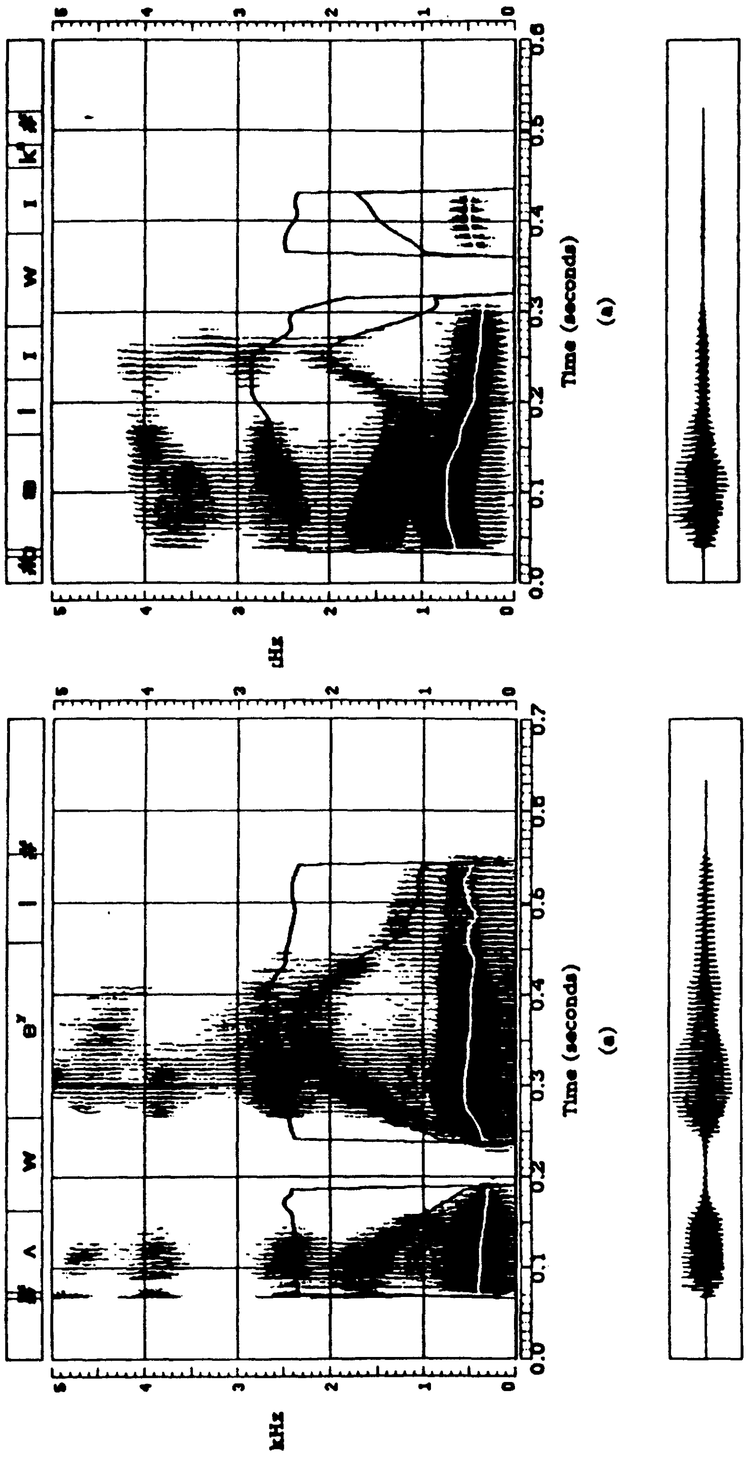


Figure 5.5: The /w/'s in the words "bewail" and "bailiwick" are omitted from the detected voiced sonorant regions. (a) Wide band spectrogram with formant tracks overlaid. (b) Waveform.

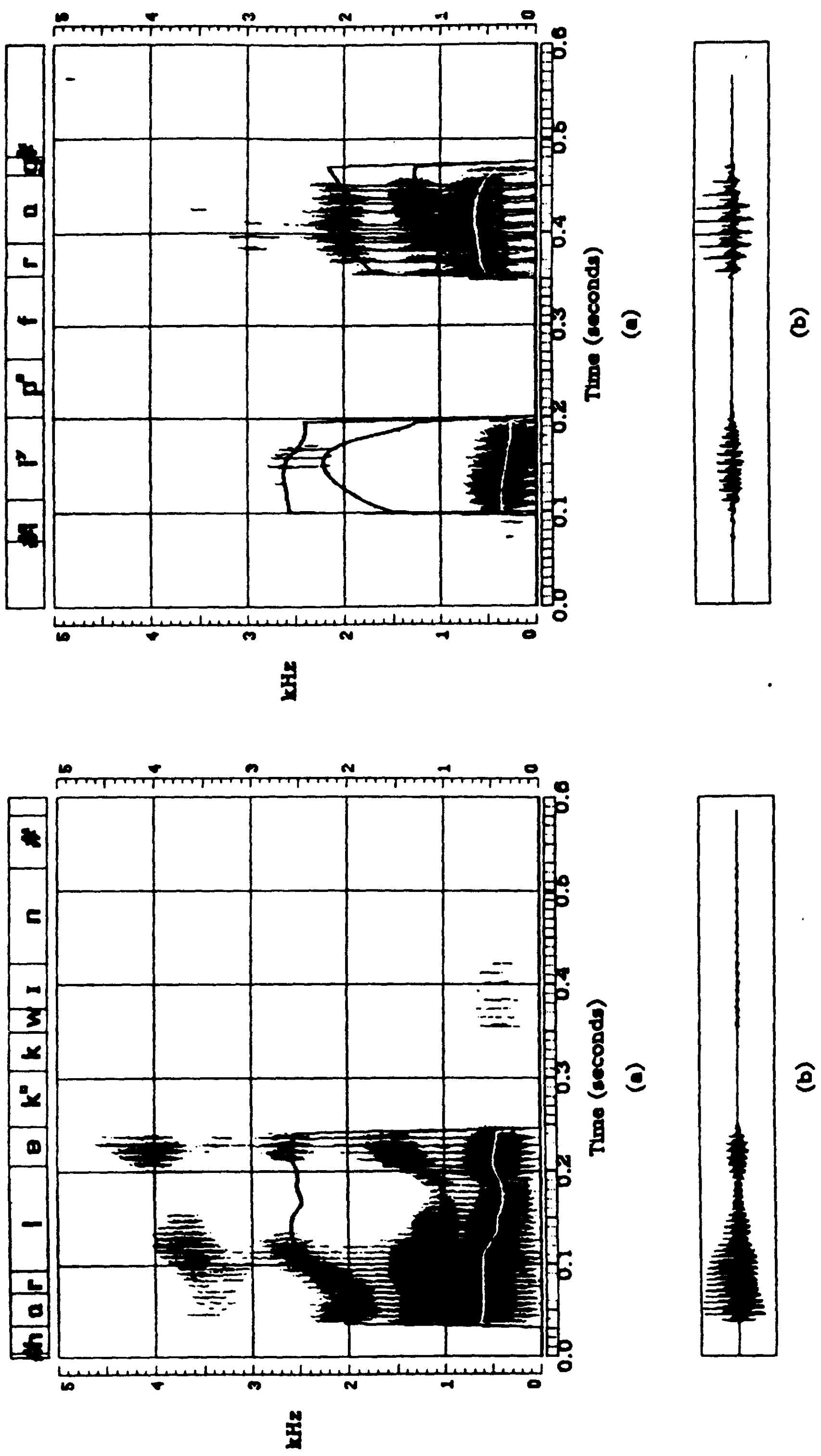


Figure 5.6: Wide band spectrogram with formant tracks overlaid of "harlequin" and "leapfrog."

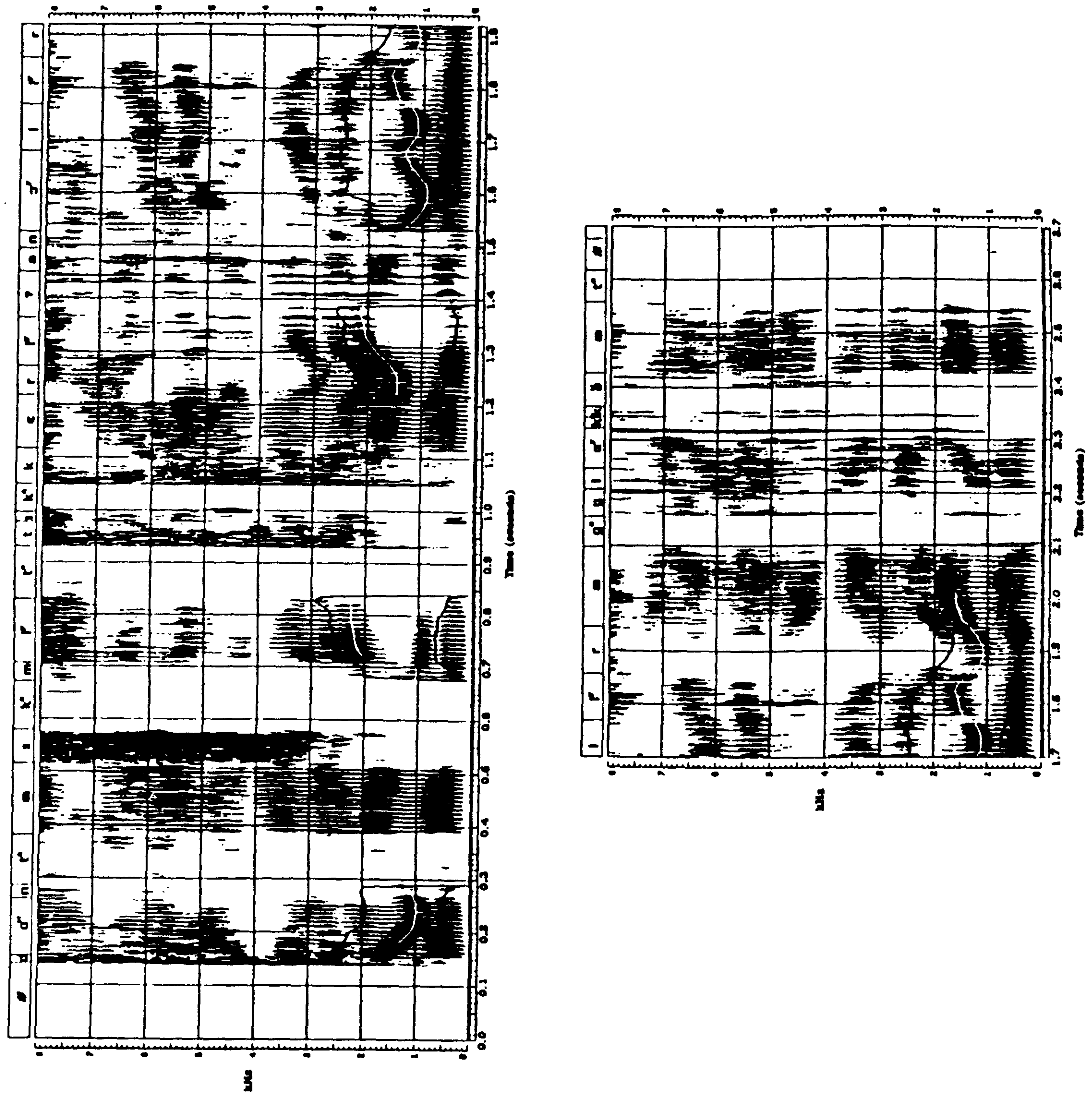
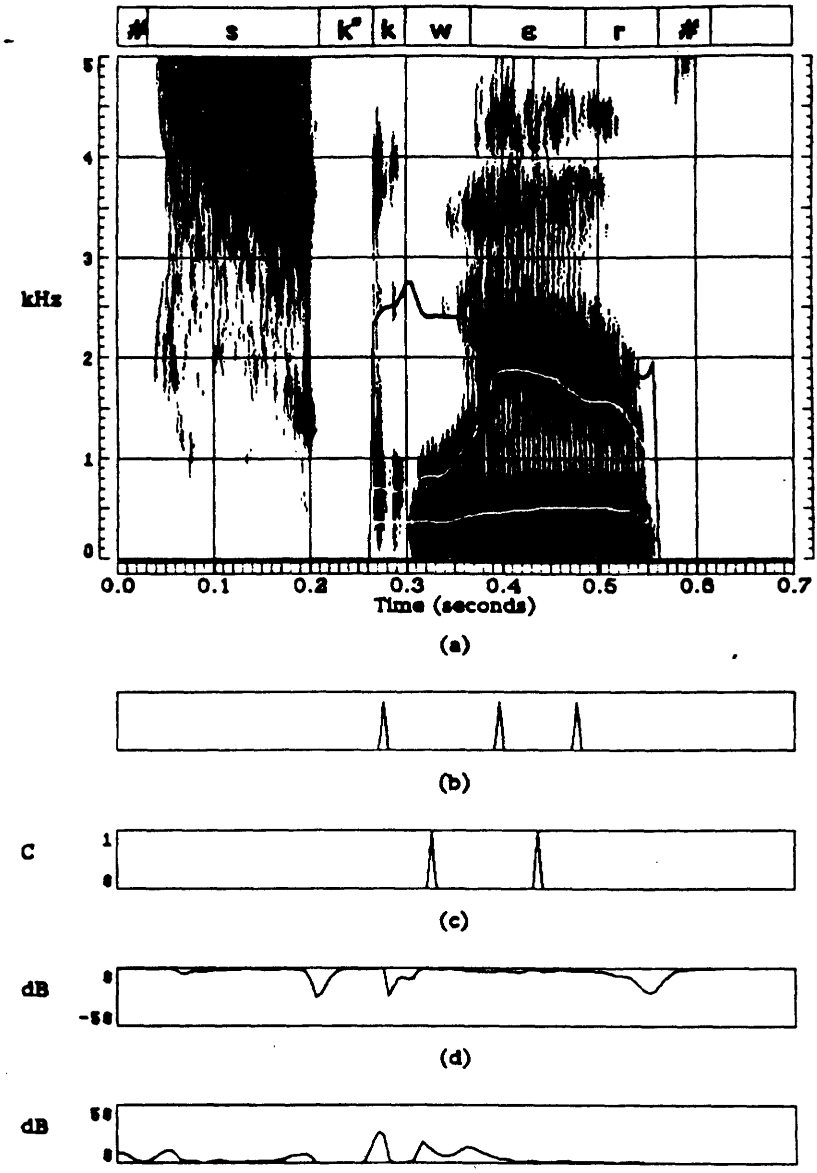


Figure 5.7: Wide band spectrogram with formant tracks overlaid of the sentence "Don't ask me to carry an oily rag like that."

/w/, and the word-initial /l/ in "leapfrog" are left out of the detected voiced sonorant regions. In addition, the word "like," which contains a word-initial /l/, in the sentence in Figure 5.7, is omitted. As in the previous examples, the semivowels in Figure 5.6 are omitted because of their relatively low amplitude. However, in the latter case, the word "like" as well as several other sounds in the sentence are excluded because of their strong high frequency energy. Although the sentences in the TI corpus were recorded with a close-talking microphone comparable to that used in the recording of the words in Database-1 and Database-2, the placement of the microphone was different. In the recording of Database-1 and Database-2, the microphone was placed about 2 centimeters in front of the mouth. However, in the recording of Database-3, the microphone touched the mouth. As a result, the sounds in the TI corpus have considerably more high frequency energy. Thus, since the ratio of low- to high-frequency energy of the utterances in Database-3 can be considerably different from that of the other utterances used to develop the voiced sonorant detector, several voiced and sonorant sounds in this corpus were excluded from the detected sonorant regions. As for the semivowels contained in Database-3, only the /l/ shown in Figure 5.7 and a /y/ were excluded from detected voiced sonorant regions.

The problem of excluding very weak sonorant sounds can possibly be corrected in several ways. One possible correction is to adjust the relative energy threshold used to extract voiced regions. However, such a modification may result in the inclusion of stop gaps. Alternatively, estimates of the voiced and sonorant regions can be refined by tracking formants everywhere (not using continuity constraints outside of the initially detected voiced sonorant regions) and expanding the initial region to include areas where continuous tracks are extracted.

In addition to excluding a few voiced and sonorant sounds, the voiced sonorant detector also included some unvoiced and nonsonorant sounds. In some instances, such inclusions resulted in a semivowel which was not classified because its context was not correctly recognized. For example, consider the classification of the /w/ in the word "square" shown in Figure 5.8. As can be seen from the overlaid formant tracks, the low-frequency /k/ burst is included in the detected voiced sonorant region. As a result, an energy dip, shown in part c, is placed in the beginning of the /w/ and an energy peak, shown in part b, occurs within the /k/. Therefore, the prevocalic /w/ is considered to be intervocalic. As a result, it is analyzed by the intersonorant classification strategy. While this energy dip region has most of the features for an



(a)

(b)

(c)

(d)

(e)

Figure 5.8: An illustration of some acoustic events marked in the word "square." (a) Wide band spectrogram with formant tracks overlaid. (b) Location of energy peaks. (c) Location and confidence of energy dips. (c) Offset waveform. (d) Onset waveform.

intervocalic /w/, the offset occurring at approximately 280 msec is too abrupt for a /w/ in this context. Although this offset is due to the /k/ burst, it is taken to be the offset of a preceding vowel. Thus, the /w/ is not classified.

This may be a difficult problem to solve without a reliable pitch detector. On the other hand, some modifications in the voiced and/or sonorant parameters may be sufficient. For example, changing the voiced parameter from a bandlimited energy from 0 Hz to 700 Hz to one from 0 Hz to 300 Hz and using a similar relative measure (the threshold may need to be changed) should exclude many of the low-frequency stop bursts from the detected voiced region. In addition, a change in the sonorant parameter may also give better results. That is, it may be more appropriate to look at only low-frequency energy as opposed to a ratio of low-frequency energy and high-frequency energy.

Finally, intersonorant energy dips are sometimes detected in vowels and in semivowels which are prevocalic or postvocalic. Unlike the case just discussed, these intersonorant energy dips are not due to errors in the voiced sonorant detector. Such energy dips sometimes cause semivowels to go undetected or to be analyzed by an inappropriate algorithm which results in their being unclassified. Examples of this phenomenon are shown in Figure 5.9. In the word "prime," shown on the left side, an energy dip occurs during the /r/. As a result, an energy peak is placed at the beginning of the /r/. Consequently, the upward movement in F3 from the /r/ and through the /a^y/ is not detected by the sonorant-initial F3 dip detector. (Recall from the discussion of Section 4.3.1. that the detection of significant formant movement in sonorant-initial semivowels is dependent upon accurate detection of the first vowel region which is assumed to occur around the first energy peak in the detected voiced sonorant region.) Instead, the /r/ is analyzed by the intersonorant recognition algorithm. While the dip region has all of the features for an /r/, the movement in F3 is not appropriate for an intervocalic /r/. Instead of F3 increasing slightly before the energy dip, F3 should decrease from its value within the preceding vowel if the /r/ is indeed intervocalic. As a consequence, the /r/ is not classified.

A similar situation occurs for the /r/ in "cartwheel." Due to the placement of the energy dip and energy peaks in the first voiced sonorant region, the sonorant-final F3 dip detector does not mark the downward movement in F3. (Again, detection of significant formant movement signalling the presence of sonorant-final semivowels depends upon the accurate detection of the last vowel region which is assumed to occur

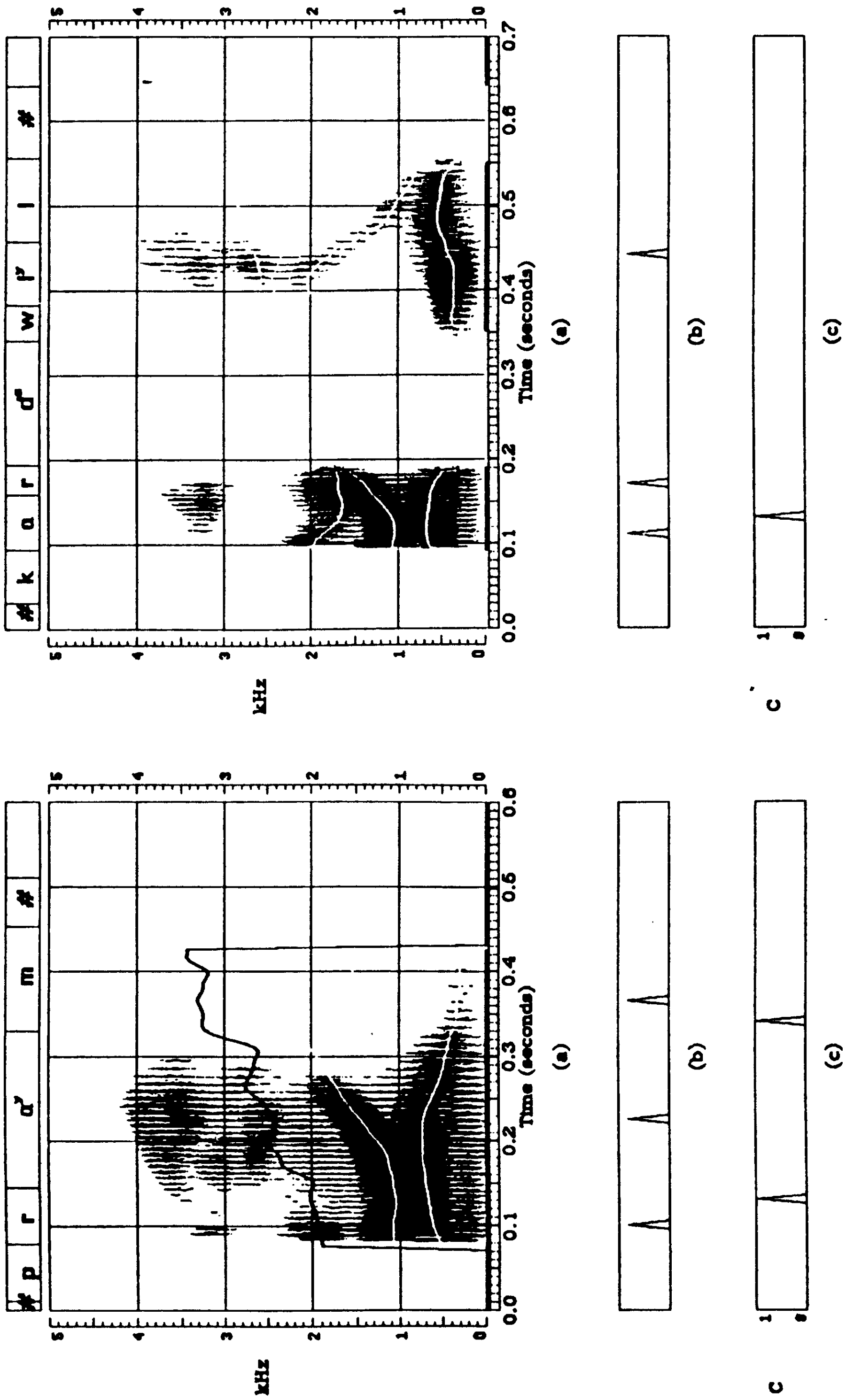


Figure 5.9: An illustration of some acoustic events marked in the words "prime" and "cartwheel." (a) Wide band spectrogram with formant tracks overlaid. (b) Location of energy peaks. (c) Location and confidence of energy dips.

around the last energy peak within the detected voiced sonorant region.) Instead, the end of the transcribed /ɑ/ is analyzed by the intersonorant recognition strategy. As in the previous example, the dip region has the necessary features for an /r/ classification, but the nearly flat F3 between the energy dip and the end of the voiced sonorant region is not appropriate for an intervocalic /r/. Thus, the /r/ is not classified.

5.5 Semivowel Recognition Results

The overall recognition results for the data bases are compared in Table 5.1. On the left side of the table are the detection results which are given separately for each data base. The top row specifies the semivowel tokens as transcribed. The following rows show the actual number of semivowels that were transcribed (# tokens), the percentage of semivowels detected by one or more acoustic event (detected), and the percentage of semivowels detected by each type of acoustic event marked by the detection algorithms. For example, the detection table for Database-1 states that 97% of the transcribed /w/'s contained an F2 dip within their segmented region.

The classification results for each data base are given on the right side of the table. As before, the top row specifies the semivowel tokens as transcribed. The following rows show the number of semivowel tokens transcribed, the number which were undetected (this number is the complement of the percent detected given in the detection results) and the percentage of those semivowel tokens transcribed which were classified by the semivowel rules. For example, the results for Database-1 show that 90% of the 558 tokens of /r/'s which were transcribed were correctly classified. The term "nc" (in the bottom row) means that one or more semivowel rules was applied to the detected sound, but the classification score(s) was less than 0.5.

Recall from the discussion in Section 5.2 that there will not always be agreement between the detection and classification results. That is, a semivowel which is considered undetected may show up in the classification results as being recognized. Thus,the numbers in a column within the classification results may not always add up to 100%.

The teased recognition results are given in Tables 5.2 - 5.7 (see pages 173 - 178). Included in the tables are the classification results for nasals. These results are given because the nasals are the only other consonants which are sonorant in all contexts. In addition, as mentioned in Chapter 3, a parameter which captures the feature *nasal* is not included in the recognition system. Thus, we expect there to be some misclas-

Table 5.1: Overall Recognition Results for the Semivowels.

Detection **Classification**

Database-1

	w	l	r	y
# tokens	369	540	558	222
detected(%)	98.6	96.7	97.4	96.2
Energy dip(%)	47	51	36	35
F2 dip(%)	97	83	46	0
F2 peak(%)	0	0	1	92
F3 dip(%)	41	10	95	2
F3 peak(%)	21	54	1	78

	w	l	r	y
# tokens	369	540	558	222
undetected(%)	1.4	3.3	2.6	2.9
w(%)	52	7.5	3.4	0
l(%)	9.1	55.7	0	0
w-l(%)	31.4	30.4	0	0
r(%)	4	.2	90	0
y(%)	0	0	0	93.7
nc(%)	2	3	4.7	4.9

Database-2

	w	l	r	y
# tokens	181	274	279	105
detected(%)	98.3	98.5	96.4	98.1
Energy dip(%)	49	59	44	41
F2 dip(%)	93	85	49	0
F2 peak(%)	0	0	1	95
F3 dip(%)	37	7	90	0
F3 peak(%)	30	69	2	87

	w	l	r	y
# tokens	181	274	279	105
undetected(%)	1.7	1.5	4.3	2.8
w(%)	48	3.6	1.9	0
l(%)	12.7	57.7	0	0
w-l(%)	29	33.8	0	0
r(%)	3.5	.4	91.3	0
y(%)	0	0	0	84.9
nc(%)	6.7	2.9	4.3	13.3

Database-3

	w	l	r	y
# tokens	28	40	49	23
detected(%)	96.4	92.5	100	96
Energy dip(%)	61	89	61	57
F2 dip(%)	93	83	65	0
F2 peak(%)	0	0	0	91
F3 dip(%)	47	36	94	0
F3 peak(%)	61	50	4	70

	w	l	r	y
# tokens	28	40	49	23
undetected(%)	3.6	7.5	0	4
w(%)	46	10	0	0
l(%)	21.6	52.6	0	0
w-l(%)	21.6	24.7	0	0
r(%)	7.1	0	89.8	0
y(%)	0	0	0	78.5
nc(%)	0	5.1	10.2	17.2

sifications of nasals as semivowels.

Note that detection results are not given for the nasals. While formant dips and peaks are marked in their hand-transcribed regions, it is not clear how to interpret these results since the formants are influenced by the presence of nasal poles and zeros. The nasals detected by energy dips can be inferred from the undetected results given in the classification tables.

As can be seen in Tables 5.2 - 5.4, the sonorant-initial semivowels are divided into the classes: semivowels which are not preceded by a consonant, semivowels which are preceded by a voiced consonant, and semivowels which are preceded by an unvoiced consonant. In the latter two categories, the semivowel may or may not be in the same syllable as the preceding consonant. Thus, the category for semivowels which follow an unvoiced consonant contains both of the /r/'s in the words "misrule" and "enshrine."

The intersonorant semivowels which are given in Tables 5.5 and 5.6 are separated on the basis of whether the semivowels are intervocalic or in a cluster with either another semivowel or a nasal. The latter division includes both the /y/ in "granular" where the intersonorant /y/ occurs in an intervocalic sonorant consonant cluster and the /r/ in "snarl" where the intersonorant /r/ occurs in a word-final sonorant consonant cluster.

Recall that the acoustic study of Chapter 3 shows that typically nonsonorant and voiced consonants may appear to be sonorant when they occur between two sonorant sounds. Thus, some voiced consonant and semivowel clusters such as the /v/ and /r/ in "everyday" are realized acoustically as an intersonorant sonorant consonant cluster. However, since this phenomenon does not always occur, results for such semivowels are given in either the data for the sonorant-initial semivowels or the data for the sonorant-final semivowels.

When comparing the recognition results of the three data bases, the many differences between Database-3 and the other corpora which were summarized in Section 2.1 should be kept in mind. In addition to these distinctions, the sparseness of the semivowels in Database-3 affects the recognition results. As can be seen from the teased results, no /y/'s occur in intervocalic position and and all prevocalic semivowels are preceded by a consonant. In addition, only /r/'s which are not syllable-final occur in sonorant-final position. Thus, several semivowels in particular contexts in Database-1 and Database-2 that receive high recognition scores are not covered in Database-3.

In view of the differences between the data bases, the detection and classification results are fairly consistent. In terms of detection, the results from all three data bases show the importance of using formant information in addition to energy measures. Across contexts, F2 minima are most important in locating /w/'s and /l/' s, F3 minima are most important in locating /r/'s and F2 maxima are most important in locating /y/'s.

When in an intervocalic context (see Table 5.5), however, the detection results using only energy dips compare favorably with those using the cited formant dip/peak. Note that 95% of the intervocalic semivowels in Database-1 are detected by an energy dip. This is more than the 90% predicted by the acoustic study of Section 3.2.4. The reason for this difference is that, while energy dips which were less than 2 dB were not considered significant in the acoustic study, such energy dips were not disregarded in the recognition system if a formant dip and/or peak also occurred in the dip region marked by the surrounding energy peaks.

There are a few events listed in the detection results which, at first glance, appear strange. In each data base, some of the /r/'s contained an F3 peak in addition to an F3 dip. However, in all of these instances, the /r/ was adjacent to a coronal consonant such as the /r/ which precedes the /s/ in "foreswear" and the /r/ which precedes the /ð/ in "northward." Thus, there is a significant rise in F3 at the end of the /r/. Examples of this type of contextual influence are shown in Figure 5.10.

Similarly, there are a few /y/'s which, in addition to an F3 peak, contain an F3 dip. As can be seen in the words "yore," "pule" and "yon" shown in Figure 5.11, F3 starts from a value between 2500 Hz and 3000 Hz in the beginning of the /y/, and then dips to a frequency between 2000 Hz and 2400 Hz before it rises to the necessary frequency for the following sound(s) (note that an F3 dip was not marked in the /y/ of "yon" because the minimum occurred around 2400 Hz which is too high a frequency for it to be due to an /r/). This type of F3 movement was seen across all speakers in many such words. However, this finding is not reflected in the results for Database-2 and Database-3 since the F3 dip was said to occur in the hand-transcribed region of the following vowel. This phenomenon for /y/ has also been noted by Lehiste (1962) who states that this type of F3 transition is part of the phonetic distinctiveness of /y/. From her acoustic study of word-initial /y/'s, Lehiste found that the F3 transition from the /y/ into the following vowel involved a downward movement to a specified value near 2000 Hz and then a rapid movement to the target for the following vowel, if the vowel target was different from approximately 2000 Hz.

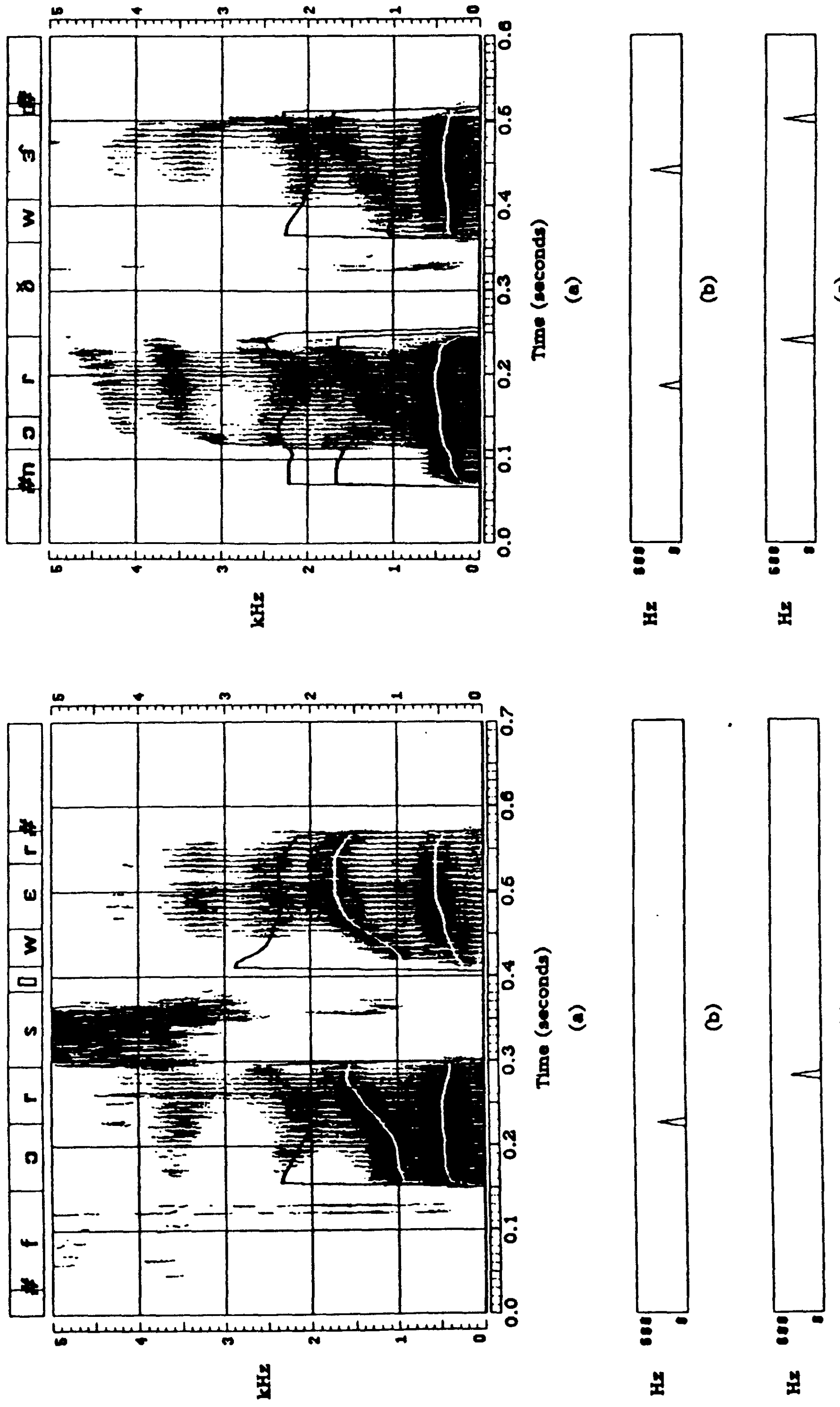


Figure 5.10: An illustration of formant movement between /r/'s and adjacent coronal consonants in the words "foreswear" and "northward." (a) Wide band spectrogram with formant tracks overlaid. (b) Location and depth of F3 dips placed by intersonorant dip detector. (c) Location and depth of F3 peaks placed by sonorant-final dip detector.

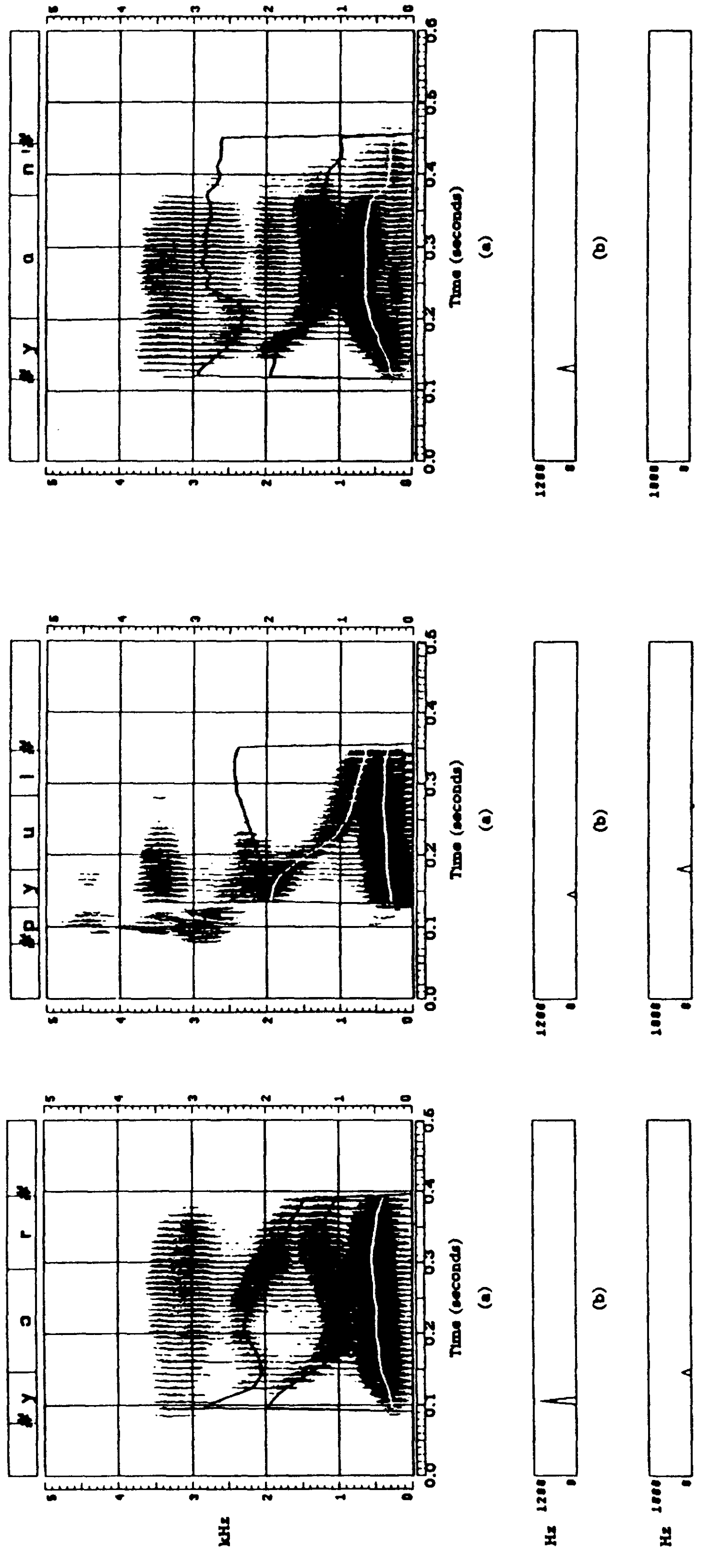


Figure 5.11: An illustration of formant movement between the /y/'s in "your," "pule" and "yon" and the following vowels. (a) Wide band spectrogram with formant tracks overlaid. (b) Location and depth of F3 peaks marked by sonorant-initial dip detector. (c) Location and depth of F3 dips marked by intersonorant dip detector.

As for the classification results, there is a considerable number of the /w/'s and /l/'s which get classified as w-l in all three data bases. This result is not surprising given the acoustic similarity of these two sounds. As the acoustic study discussed in Chapter 3 shows, no one measure used in the recognition system provides a good separation between these sounds. Note, however, that in several contexts, the system is able to correctly classify these sounds at a rate better than chance. Considering the contexts in which both sounds occur, the best results are obtained when they are word-initial (that is, sonorant-initial with no preceding consonant). As can be seen in Table 5.2, only a few /w/'s are called /l/ and only a few /l/'s are called /w/. This result is not surprising. The prevocalic /l/ allophone occurs in this context. Therefore, an abrupt spectral change due to the release of the tongue tip will usually occur between the /l/ and the following vowel. Between a /w/ and adjacent vowel(s), however, the spectral change is usually gradual. Furthermore, since there is no influence of a preceding sound, many of the sonorant-initial /w/'s have a high degree of the feature *back* and, therefore, a very low F2, whereas most of the prevocalic /l/'s will not have such a close spacing between F1 and F2. As can be seen from the other tables, the number of confusions as well as the number called /w-l/ increases significantly when they are preceded by other sounds.

If we consider the classification of /w/'s as either /w/, /l/ or /w-l/ to be correct, then the scores for the /w/'s in Database-1, Database-2 and Database-3 are 92.5%, 89.7% and 89.2%, respectively. Similarly, the lumped scores for the /l/'s in Database-1, Database-2 and Database-3 are 93.6%, 95.1% and 87.3%, respectively. Alternatively, since it is equally likely that a sound classified as /w-l/ is a /w/ or an /l/, we can assign half of the /w-l/ score to the scores for /w/ and /l/. With this tabulation, the scores for the /w/'s in Database-1, Database-2 and Database-3 are 68%, 62.5% and 56.8%, respectively; and the scores for the /l/'s in Database-1, Database-2 and Database-3 are 70.9%, 74.6% and 64.9%, respectively.

From a comparison of the results for Database-1 and Database-2, we see that a considerably larger percentage of the /w/'s in Database-2 were not classified. This result accounts for the difference in correct classification scores. Most of these "no classifications" are due to a particular speaker who had strong low frequency /k/ bursts which were included in the detected voiced sonorant region. An example of a no classification caused by the inclusion of such sounds within the voiced sonorant regions was discussed in the previous section.

The /w/ and /l/ scores for Database-3 are lowest. However, as stated earlier, some contexts occurring in Database-1 and Database-2 were not covered by Database-3. For those contexts in which /w/ and/or /l/ occur, their scores in Database-3 are comparable and sometimes better than those contained in the other data bases; however, the classification scores in the other contexts tend to be higher. Thus, it is the lack of coverage which accounts for the apparent decrease in correct recognition of these sounds and the apparent increase in the number of confusions between them.

The overall results for the /r/'s in the data bases are comparable. However, the detection and classification results for the sonorant-final /r/'s given in Table 5.7 appear to be significantly worse for Database-3. This is so because all of the sonorant-final /r/'s in Database-3 were followed by the consonant /k/ in "dark." Only 12 of the 14 repetitions of this word were transcribed with an /r/. In three of the 12 repetitions, a situation similar to that discussed for "cartwheel" in the previous section occurred. That is, an intersonorant energy dip occurred somewhere in the /ɑ/ and /r/ regions. As a result, any downward movement in F3 between the coronal consonant /d/ and the retroflexed /ɑ/, was not detected. This outcome is apparent from the detection results which state that only 75% of the /r/'s contained an F3 dip. Thus, we feel that had this data base contained some syllable-final /r/'s which were also sonorant-final, the classification score for the /r/ in this context would be comparable to that obtained for the other data bases. A finding in support of this claim is the many /ɝ/'s and /ɚ/'s contained in Database-3 which were called /r/. These syllabic /r/'s occurred in the words "your" and "water." The word "your" was also contained in Database-1 and Database-2 (in these data bases, it was spell as "yore"). However, in these data bases, this word was always transcribed with a vowel followed by an /r/.

As for the /y/'s, the overall results show that the classification scores for Database-2 and Database-3 are lower than the scores for Database-1. For Database-2, this lower score is due mainly to one of the two speakers for whom the classification of intersonorant /y/'s in clusters with nasals was poor (see Table 5.6). The reason for this poor classification is illustrated in Figure 5.12. Given on the left side are several displays corresponding to the word "banyan" which is a part of Database-1. The pattern of events illustrated is typical for the intervocalic nasal-semivowel clusters seen in this data base. In contrast, the same displays are shown for the same word said by the speaker of Database-2. The main differences between the pattern of events for these two words lies in the energy dip region which is defined by the offset preceding

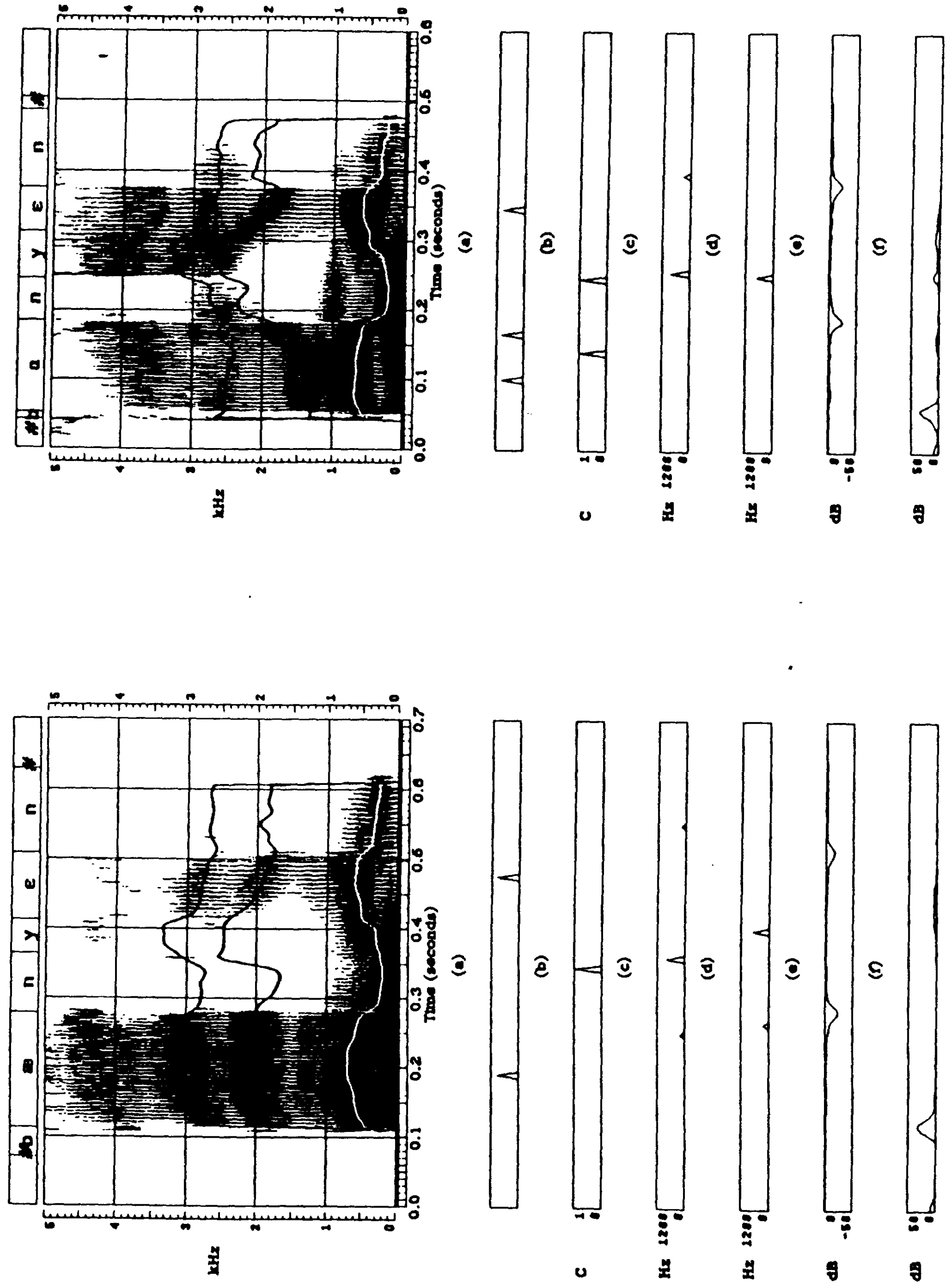


Figure 5.12: A comparison of the /ny/ regions in the words "banyan" spoken by two different speakers. (a) Wide band spectrogram with formant tracks overlaid. (b) Location of energy peaks (c) Location and confidence of energy dips. (d) Location and depth of F2 peaks. (e) Location and depth of F3 peaks (f) Offset waveform. (e) Onset waveform.

Table 5.2: Recognition Results for Sonorant-Initial Semivowels Not Adjacent to a Consonant.

Detection **Classification**

Database-1

	w	l	r	y
# tokens	70	40	56	46
detected(%)	100	90	100	95.7
Energy dip(%)	43	33	25	43
F2 dip(%)	97	70	79	0
F2 peak(%)	0	0	0	95
F3 dip(%)	57	10	98	2
F3 peak(%)	37	55	0	95

	w	l	r	y	nasal
# tokens	70	40	56	46	64
undetected(%)	0	10	0	4.3	14
w(%)	80	5	5	0	5
l(%)	1.4	63	0	0	20
w-l(%)	17.1	15	0	0	3
r(%)	0	0	95	0	5
y(%)	0	0	0	91.4	9
nc(%)	1.4	7	0	4.3	44

Database-2

	w	l	r	y
# tokens	33	21	27	19
detected(%)	97	95	96	100
Energy dip(%)	18	50	30	10
F2 dip(%)	97	81	89	0
F2 peak(%)	0	0	0	100
F3 dip(%)	36	14	93	0
F3 peak(%)	48	62	0	100

	w	l	r	y	nasal
# tokens	33	21	27	19	28
undetected(%)	3	5	4	0	7
w(%)	67	0	7.6	0	0
l(%)	9	76	3.8	0	10.7
w-l(%)	21	5	0	0	3.6
r(%)	0	0	81	0	10.7
y(%)	0	0	0	94	3.6
nc(%)	0	14	7.6	6	64.3

Table 5.3: Recognition Results for Sonorant-Initial Semivowels Adjacent to Voiced Consonants.

Detection **Classification**

Database-1

	w	l	r	y
# tokens	35	29	67	30
detected(%)	94	100	94	97
Energy dip(%)	40	55	18	13
F2 dip(%)	94	86	55	0
F2 peak(%)	0	0	1	87
F3 dip(%)	40	31	90	3
F3 peak(%)	43	48	0	77

	w	l	r	y	nasal
# tokens	35	29	67	30	0
undetected(%)	6	0	6	3	0
w(%)	37	24	6	0	0
l(%)	11	28	0	0	0
w-l(%)	40	48	0	0	0
r(%)	3	0	88	0	0
y(%)	0	0	0	90	0
nc(%)	3	0	0	7	0

Database-2

	w	l	r	y
# tokens	18	13	31	14
detected(%)	100	100	100	100
Energy dip(%)	56	62	42	21
F2 dip(%)	94	100	52	0
F2 peak(%)	0	0	0	79
F3 dip(%)	61	8	94	0
F3 peak(%)	17	62	6	93

	w	l	r	y	nasal
# tokens	18	13	31	14	0
undetected(%)	0	0	0	0	0
w(%)	78	8	0	0	0
l(%)	0	38	0	0	0
w-l(%)	22	54	0	0	0
r(%)	0	0	97	0	0
y(%)	0	0	0	79	0
nc(%)	0	0	3	21	0

Database-3

	w	l	r	y
# tokens	0	13	13	9
detected(%)	0	92	100	89
Energy dip(%)	0	31	31	0
F2 dip(%)	0	85	85	0
F2 peak(%)	0	0	0	77
F3 dip(%)	0	23	100	0
F3 peak(%)	0	38	0	55

	w	l	r	y	nasal
# tokens	0	13	13	9	0
undetected(%)	0	8	0	11	0
w(%)	0	0	0	0	0
l(%)	0	46	0	0	0
w-l(%)	0	38	0	0	0
r(%)	0	0	92	0	0
y(%)	0	0	0	56	0
nc(%)	0	8	8	33	0

Table 5.4: Recognition Results for Sonorant-Initial Semivowels Adjacent to Unvoiced Consonants.

Detection **Classification**

Database-1

	w	l	r	y
# tokens	144	123	129	69
detected(%)	98	93	98.4	97
Energy dip(%)	10	11	10	4
F2 dip(%)	94	85	53	0
F2 peak(%)	0	0	3	94
F3 dip(%)	53	27	95	3
F3 peak(%)	19	46	1	72

	w	l	r	y	nasal
# tokens	144	123	129	69	4
undetected(%)	2	7	1.6	0	25
w(%)	51	20	4.6	0	0
l(%)	11	32	.8	0	25
w-l(%)	25	37	0	0	0
r(%)	8	1	83.7	0	0
y(%)	0	0	0	90	25
nc(%)	3	3	9.3	10	25

Database-2

	w	l	r	y
# tokens	69	56	60	30
detected(%)	97	100	93	100
Energy dip(%)	26	16	10	13
F2 dip(%)	87	93	58	0
F2 peak(%)	0	0	3	93
F3 dip(%)	45	20	85	0
F3 peak(%)	22	71	0	87

	w	l	r	y	nasal
# tokens	69	56	60	30	2
undetected(%)	3	0	7	0	0
w(%)	45	12.5	2	0	0
l(%)	16	25	0	0	0
w-l(%)	22	62.5	0	0	50
r(%)	3	0	86	0	0
y(%)	0	0	0	97	0
nc(%)	12	0	5	3	50

Database-3

	w	l	r	y
# tokens	14	0	0	0
detected(%)	92.86	0	0	0
Energy dip(%)	21	0	0	0
F2 dip(%)	86	0	0	0
F2 peak(%)	0	0	0	0
F3 dip(%)	57	0	0	0
F3 peak(%)	64	0	0	0

	w	l	r	y	nasal
# tokens	14	0	0	0	13
undetected(%)	7.14	0	0	0	23
w(%)	71.43	0	0	0	38
l(%)	7.14	0	0	0	23
w-l(%)	7.14	0	0	0	0
r(%)	7.14	0	0	0	8
y(%)	0	0	0	0	0
nc(%)	0	0	0	0	8

Table 5.5: Recognition Results for Intervocalic Semivowels.

Detection **Classification**

Database-1

	w	l	r	y
# tokens	73	188	145	44
detected(%)	100	100	100	98
Energy dip(%)	99	97	93	86
F2 dip(%)	100	88	52	0
F2 peak(%)	0	0	0	95
F3 dip(%)	23	2	99	0
F3 peak(%)	16	43	0	89

	w	l	r	y	nasal
# tokens	73	188	145	44	88
undetected(%)	0	0	0	2	0
w(%)	35	1	3	0	2
l(%)	14	54	0	0	24
w-l(%)	48	43	0	0	1
r(%)	3	0	97	0	6
y(%)	0	0	0	100	14
nc(%)	0	2	0	0	53

Database-2

	w	l	r	y
# tokens	42	99	79	25
detected(%)	100	100	100	96
Energy dip(%)	88	96	96	84
F2 dip(%)	100	86	53	0
F2 peak(%)	0	0	0	96
F3 dip(%)	22	1	96	0
F3 peak(%)	37	57	0	72

	w	l	r	y	nasal
# tokens	42	99	79	24	42
undetected(%)	0	0	0	4	0
w(%)	21	1	1.25	0	16.6
l(%)	19	57	0	0	4.7
w-l(%)	48	40	0	0	4.7
r(%)	10	1	97.5	0	4.7
y(%)	0	0	0	87.5	4.7
nc(%)	2	1	1.25	12.5	69

Database-3

	w	l	r	y
# tokens	14	13	24	0
detected(%)	100	100	100	0
Energy dip(%)	100	92	96	0
F2 dip(%)	100	70	83	0
F2 peak(%)	0	0	0	0
F3 dip(%)	36	15	100	0
F3 peak(%)	57	54	0	0

	w	l	r	y	nasal
# tokens	14	13	24	0	8
undetected(%)	0	0	0	0	37.5
w(%)	21	0	0	0	0
l(%)	36	62	0	0	0
r(%)	7	0	96	0	12.5
y(%)	0	0	0	0	0
nc(%)	0	0	4	0	25

Table 5.6: Recognition Results for Semivowels in Intersonorant Cluster.

Detection **Classification**

Database-1

	w	l	r	y
# tokens	47	57	73	33
detected(%)	100	93	92	92
Energy dip(%)	89	62	23	38
F2 dip(%)	100	52	18	0
F2 peak(%)	0	0	0	85
F3 dip(%)	11	4	90	0
F3 peak(%)	21	50	0	54

	w	l	r	y	nasal
# tokens	47	57	73	33	48
undetected(%)	0	7	8	8	6
w(%)	51	9	0	0	0
l(%)	6	47	0	0	8
w-l(%)	40	30	0	0	4
r(%)	0	0	85	0	0
y(%)	0	0	0	100	4
nc(%)	2	7	12	0	78

Database-2

	w	l	r	y
# tokens	19	32	36	18
detected(%)	100	90.6	92	89
Energy dip(%)	95	56	39	71
F2 dip(%)	95	56	19	0
F2 peak(%)	0	0	0	100
F3 dip(%)	21	9	86	0
F3 peak(%)	21	78	0	82

	w	l	r	y	nasal
# tokens	19	32	36	18	26
undetected(%)	0	9.4	11	11	4
w(%)	58	3	0	0	15.4
l(%)	5	59.4	0	0	11.5
w-l(%)	32	22	3	0	3.8
r(%)	0	0	86	0	7.7
y(%)	0	0	0	56	3.8
nc(%)	5	6.2	11	33	53.8

Database-3

	w	l	r	y
# tokens	0	14	0	14
detected(%)	0	86	0	100
Energy dip(%)	0	64	0	93
F2 dip(%)	0	21	0	0
F2 peak(%)	0	0	0	100
F3 dip(%)	0	36	0	0
F3 peak(%)	0	14	0	79

	w	l	r	y	nasal
# tokens	0	14	0	14	0
undetected(%)	0	14	0	0	0
w(%)	0	29	0	0	0
l(%)	0	50	0	0	0
w-l(%)	0	0	0	0	0
r(%)	0	0	0	0	0
y(%)	0	0	0	93	0
nc(%)	0	7	0	7	0

Table 5.7: Recognition Results for Sonorant-Final Semivowels.

Detection **Classification**

Database-1

	l	r
# tokens	103	88
detected(%)	99	97
Energy dip(%)	8	10
F2 dip(%)	93	18
F2 peak(%)	0	1
F3 dip(%)	1	95
F3 peak(%)	89	7

	l	r	nasal
# tokens	103	88	260
undetected(%)	1	3	37
w(%)	0	2	0
l(%)	97	0	5
w-l(%)	1	0	3
r(%)	0	91	1
y(%)	0	0	3
nc(%)	1	6	51

Database-2

	l	r
# tokens	53	48
detected(%)	100	94
Energy dip(%)	38	9
F2 dip(%)	91	26
F2 peak(%)	0	2
F3 dip(%)	0	85
F3 peak(%)	89	9

	l	r	nasal
# tokens	53	48	134
undetected(%)	0	6	40
w(%)	0	2	0
l(%)	90.6	0	6
w-l(%)	5.6	0	2
r(%)	0	90	0
y(%)	0	0	2
nc(%)	3.8	2	50

Database-3

	l	r
# tokens	0	12
detected(%)	0	100
Energy dip(%)	0	25
F2 dip(%)	0	8
F2 peak(%)	0	0
F3 dip(%)	0	75
F3 peak(%)	0	17

	l	r	nasal
# tokens	0	12	23
undetected(%)	0	0	70
w(%)	0	0	9
l(%)	0	0	4
w-l(%)	0	0	0
r(%)	0	75	0
y(%)	0	0	0
nc(%)	0	25	17

and the onset following the intersonorant energy dip occurring within the /n/. The location and confidence of the energy dip is shown in part b. In the word on the left, the offset, which can be seen in part f, occurs at about 280 msec. The onset, which can be seen in part g, occurs at about 420 msec. Thus, the duration of the energy dip region is approximately 140 msec and the region includes both the /n/ and the /y/. In the word on the right, however, the offset occurs at about 190 msec and the onset occurs at about 260 msec, so that the duration of the energy dip region is only 70 msec. In this case, the energy dip region includes only the /n/. Recall that duration is one of the main cues used to determine if an intervocalic dip region contains one or two sonorant consonants. Thus, the recognition system correctly decides that the energy dip region in the word on the left contains two sonorant consonants. Consequently, the abrupt offset marking the beginning of the /n/ is not included in the classification of the /y/. However, in the case of the energy dip region in the word on the right, the algorithm decides that it contains only one sonorant consonant. Thus, the abrupt offset due the /n/ and the F2 and F3 peaks due to the /y/ are assumed to be cues for the same sound. Consequently, this /y/, as well as most /y/'s occurring in this context spoken by this speaker, is not classified.

5.6 Consonants called Semivowels

The teased results as well as Table 5.8 show that many nasals are called semivowels. As stated earlier, one main reason for this confusion is the lack of a parameter which captures the feature *nasal.* Presently, the main cues used for the nasal-semivowel distinction are the offsets and onsets. This accounts for the generally higher misclassification of nasals as /l/. While the rate of spectral change is often abrupt between nasals and adjacent sounds, the data of Section 3.2.5 show that this is not always the case, particularly when the nasals are adjacent to unstressed vowels. Thus, they are sometimes classified as other semivowels as well.

In addition to the nasals, a few flaps, /h/'s and sonorant-like voiced consonants are also called semivowels. The latter sounds are grouped into a class called "Others" and their recognition results are shown in Table 5.8. Examples of these types of confusions are shown in Figure 5.13.

In "frivolous," the intervocalic /v/ is classified as an /l/. Note that it does have frequency values in the range of those acceptable for an /l/. In "waterproof," the F3

Table 5.8: Recognition of Other Sounds as Semivowels.

Database-1

	nasals	others	vowels
# tokens	464	508	2385
undetected(%)	24	81.5	
w(%)	1	1	1
l(%)	11	3.3	5.5
w-l(%)	3	.8	2
r(%)	2	.6	6
y(%)	6	1.4	8.6
nc(%)	53	11.4	39

Database-2

	nasals	others	vowels
# tokens	232	135	1184
undetected(%)	24	69	
w(%)	5	0	1
l(%)	7	6	5
w-l(%)	3	1	4
r(%)	3	2	4
y(%)	3	3	10
nc(%)	55	19	42

Database-3

	nasals	others	vowels
# tokens	44	121	350
undetected(%)	50	73	
w(%)	15	0	2
l(%)	13	2.5	9
w-l(%)	0	0	4
r(%)	5	2.5	15
y(%)	0	5	9
nc(%)	17	17	62

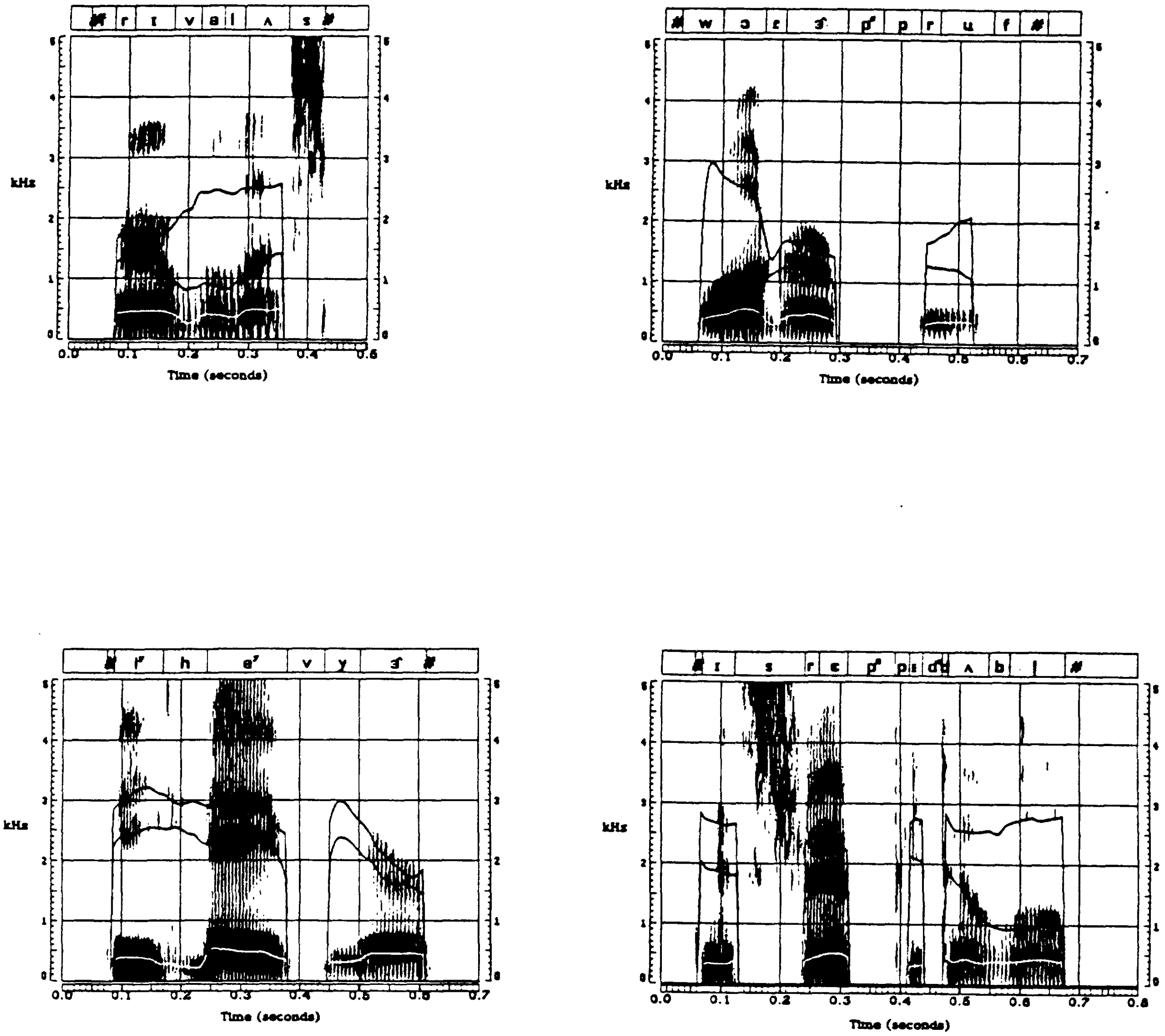


Figure 5.13: Wide band spectrograms with formant tracks overlaid of four words which contain consonants that were misclassified as semivowels. The /v/ in "frivolous" was classified as an /l/. The /ɾ/ in "waterproof" was classified as /r/. The /h/ in "behavior" was classified as /y/. The /b/ in "disreputable" was classified as /w-l/.

dip occurring in the /ɾ/ resulted in it being classified as an /r/. Recall that the /r/ rules will classify the detected sound as an /r/ if it is determined to be "retroflex" with either a "close F2 and F3" or a "maybe close F2 and F3." Since the /ɾ/ has these properties, the abrupt onset and offset surrounding it were not used in its classification.

The /h/ in "behavior" occurs after the /y/ offglide in the vowel /i/ and before another front vowel. Thus, it was probably articulated with a vocal tract configuration similar to that of a /y/. As can be seen, it has formant frequencies in the range of those acceptable for a /y/. As a result, it was misclassified as this semivowel. Finally, the /b/ in "disreputable" was classified as /w-l/. Note that, in addition to formant frequencies acceptable for a /w/ and an /l/, the /b/ does appear to be sonorant, and the rate of spectral change between it and the surrounding vowels is gradual.

In conclusion, the nonsemivowel consonants do share some of the features expected of the assigned semivowels such that the confusions made are not random. However, it is apparent that more features are needed to make the necessary distinctions. For example, the property "breathiness" may be the only additional cue needed to recognize that the /h/ in "behavior" is indeed an /h/ and not a /y/.

5.7 Vowels called Semivowels

The classification results for the vowels are also given in Table 5.8. No detection results are given for the vowels since different portions of the same vowel may be detected and labelled a semivowel. For example, across several of the speakers in Database-1 and Database-2, the beginning of the /ɔ^{y}/ in "flamboyant" was classified as either /w/, /l/ or /w-l/ and the /y/ offglide was classified as a /y/. When phenomena such as this occur, the vowel shows up in the results as being misclassified as /y/ and either /w/, /l/ or /w-l/. Similarly, though this situation never occurred for this word, if the beginning of the /ɔ^{y}/ was detected but not classified and the /y/ offglide was classified as /y/, then the vowel would show up in the results as being not classified and as being misclassified as a /y/. Thus, for these reasons, the vowel statistics for the data bases in Table 5.8 may not add up to 100%.

As can be seen in Table 5.8, there are a number of vowels or portions thereof which are classified as semivowels. Most of the misclassifications are understandable. That is, vowels or portions thereof which are called /y/ are *high* and *front*. Vowels or portions thereof which are called /w/, /l/ and /w-l/ are *back*. Finally, vowels or portions thereof which are called /r/ are either ***retroflex*** or ***round.*** A sampling of some of the vowel portions which are called semivowels is given in Appendix B.

The classification of vowels as semivowels occurs for several reasons. First, some misclassifications occur because what has been labeled as a vowel is probably a semivowel. Examples of such possible mislabelings are shown in Figure 5.14. As can be seen, the "offglides" of these vowels do in fact appear to be semivowels.

In "stalwart," the significant rise in F3 from the beginning of the /ɑ/ region resulted in the classification of the end of the transcribed /ɑ/ as an /l/. Recall from Sections 3.2.2 and 4.3.2 that this type of F3 movement is often indicative of a postvocalic /l/. Thus, while the /l/ was not included in the transcription, it was correctly recognized as /l/.

Recall from Section 3.2.4 that the /ɚ/ in "plurality" and the /i^y/ in "queer" both contain significant intravowel energy dips which suggest that parts of them are non-syllabic. In addition, there is a significant F3 minimum in the /ɚ/ and significant F2 and F3 maxima in the /i^y/. As a result, the mid portion of the /ɚ/ was classified as /r/ and the mid portion of the /i^y/ was classified as /y/. These classifications also appear to be reasonable.

Finally, the /w/ offglide of the /ɑ^w/ in "wallflower" was classified as /w/. Although an intersonorant energy dip was not detected in the /w/ offglide, an F2 dip was detected in this region. In addition to the results of Section 3.2.4, the detection results of Table 5.5 for /w/'s show that, across the data bases, intervocalic /w/'s always contain an energy dip. (Even though energy dips occur in the /w/'s which are excluded from the detected voiced sonorant regions, they are not included in the detection results. This accounts for the result in Database-2 which states that only 88% of the intervocalic /w/'s contained energy dips.) Thus, it does not appear as if a well enunciated /w/ was produced. However, whether a clear /w/ was articulated or not, the recognition of the /w/ offglide as /w/ should not be detrimental to any system which is trying to recognize this word.

Second, such misclassifications occur because a label is being assigned too early in the recognition process. That is, as we will discuss in Chapter 6, either a label should not be assigned until more information regarding context is known, or a label should perhaps not be assigned at all. Examples of such assignments are shown in Figure 5.15. In the word "forewarn," the beginning of the first /ɔ/ is called a /w/ because of the labial F2 transition and the falling F3 transition arising from the following /r/.

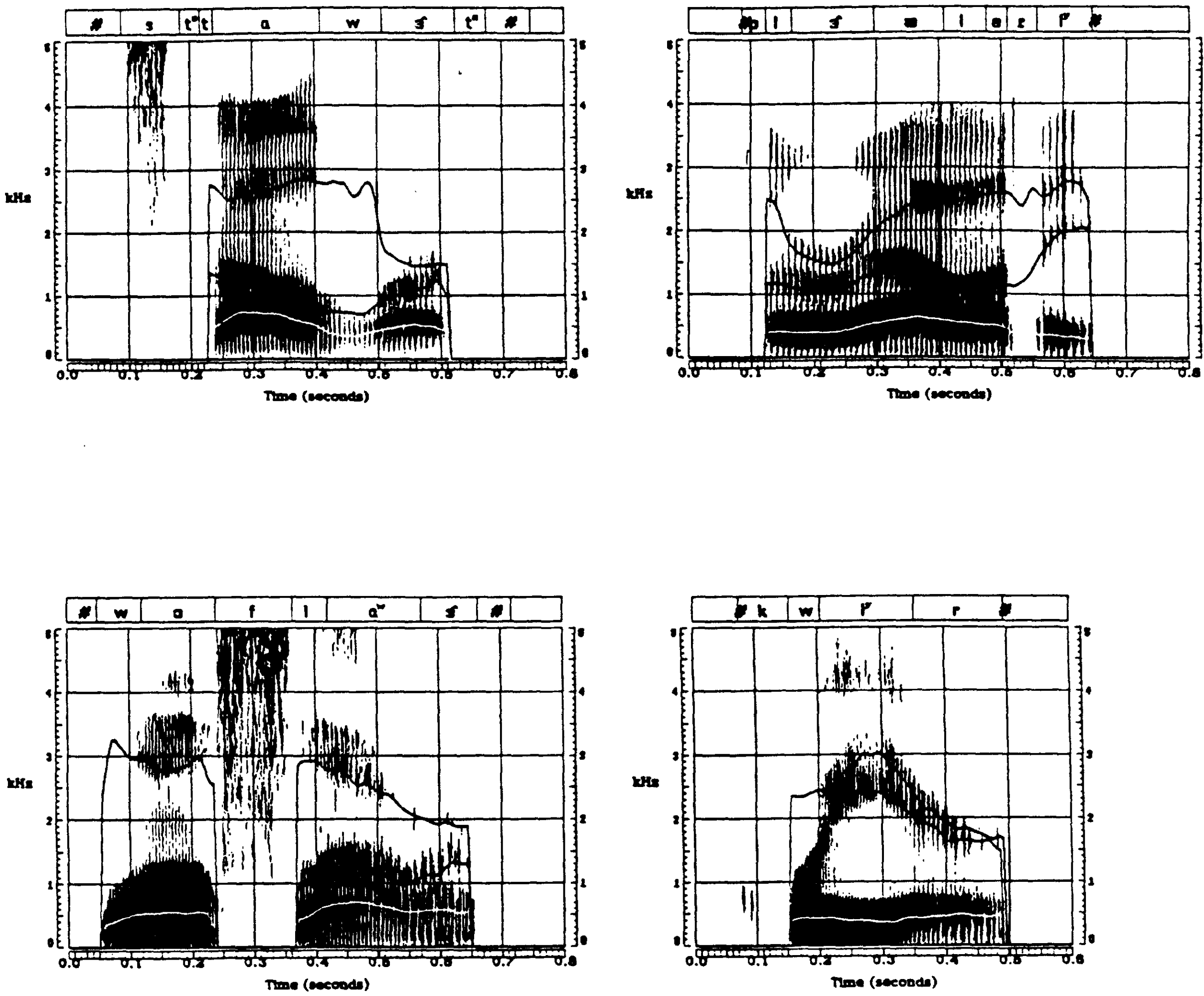


Figure 5.14: Wide band spectrograms with formant tracks overlaid of four words which contain vowels, portion of which were classified as semivowels. End of /ɑ/ in “stalwart was classified as /l/. Middle of /ɝ/ in “plurality” was classified as /r/. Middle of /i^{y}/ in “queer” was classified as /y/. End of /ɑ^{w}/ in “wallflower” was classified as /w/.

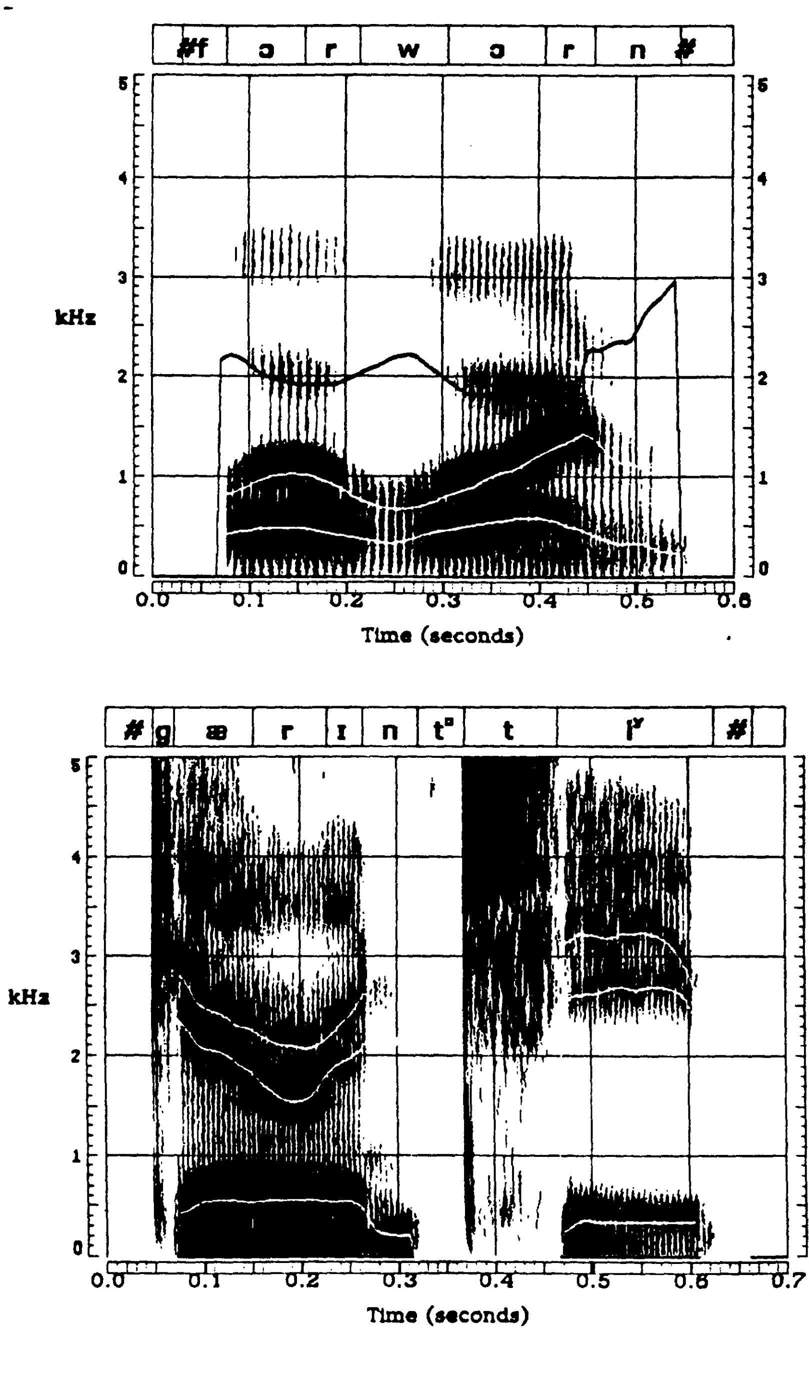


Figure 5.15: Wide band spectrograms with formant tracks overlaid of words with vowel portions which, due to contextual influence, were classified as a semivowel. Beginning of first /ɔ/ in "forewarn" was classified as /w/. Beginning of /æ/ in "guarantee" was classified as /y/.

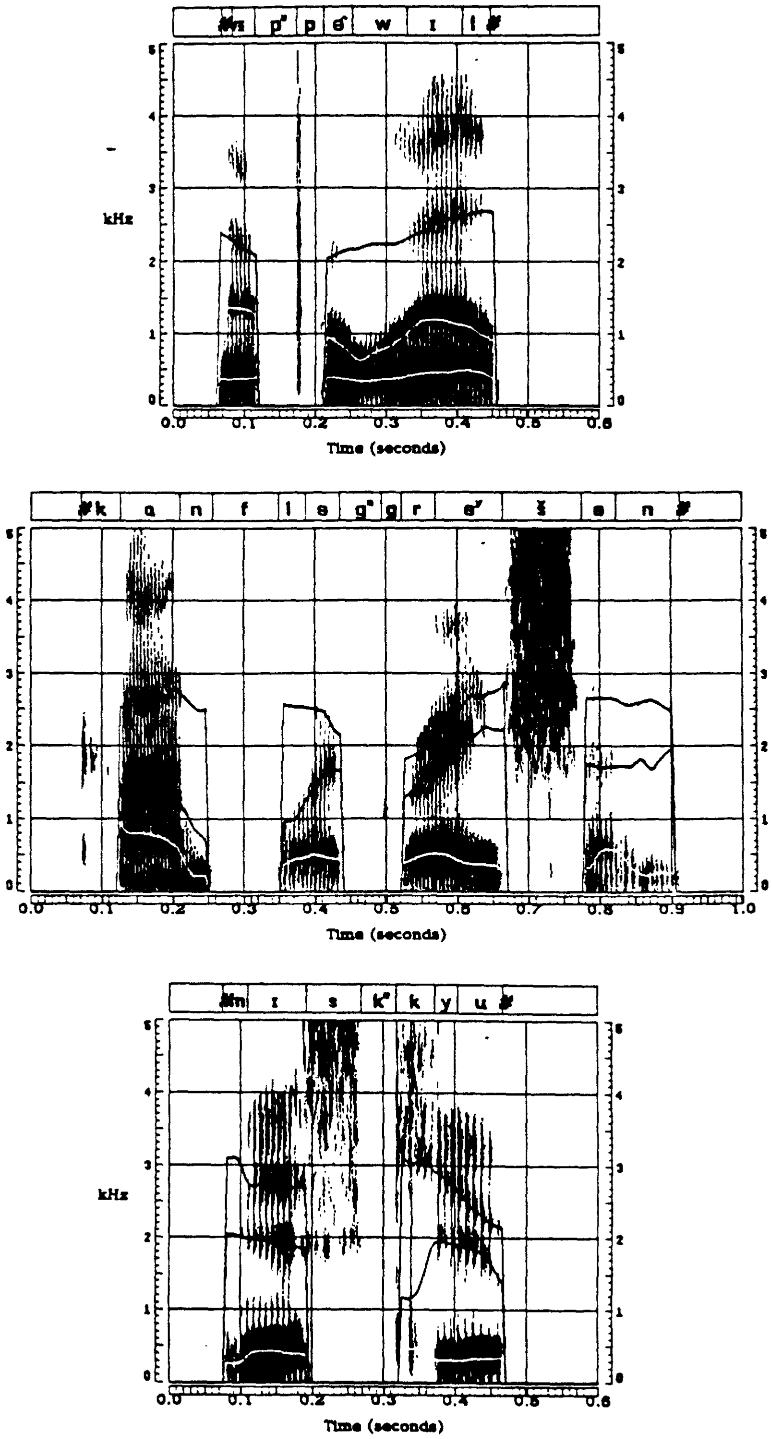


Figure 5.16: Wide band spectrograms with formant tracks overlaid of three words with vowel portions which were classified as /ɪ/. The vowel portions are the end of the /ɪ/ in "whippoorwill," the end of the first /ə/ in "conflagration" and the end of the /u/ in "miscue."

(Recall that the data of Section 3.2.2 show that /w/'s in retroflexed environments are characterized by this type of F3 movement.) Similarly, in "guarantee," the beginning of the /æ/ is called a /y/ due to the transitions of F1, F2 and F3 caused by the preceding /g/ and the following /r/. In the latter example, it is not clear that the assignment of a /y/ is incorrect since it is possible to pronounce "guarantee" with a /y/ between the /g/ and /æ/. In fact, when this utterance is played from the beginning of the sonorant region, a clear /y/ is heard.

Along these same lines are some examples shown in Figure 5.16. In the word "whippoorwill," the retroflexion due to the /ɚ/ is anticipated in the vowel /ɪ/. As can be seen, F3 falls to about 2000 Hz near the end of this vowel. This sort of spreading of the feature *retroflex* across labial consonants which do not require a particular placement of the tongue was seen for many such words in the data bases. Although it is not as clear cut, it appears as if a a similar phenomenon occurs in the word "conflagration." As before, F3 of the vowel, which in this case is the first /ə/, falls to about 2100 Hz. Presumably, the declination in F3 is due to both the /r/ which causes the /g/ burst to be low in frequency, and the /g/ which is responsible for the velar pinch in F2 and F3 of the /ə/. Finally, as mentioned earlier, some rounded vowels are called /r/. The reason that this happens is shown in the word "miscue." Although F3 typically rose during the /w/ offglide of a sonorant-final /u/, as can be seen in Figure 5.16, F2 and F3 both fall from the /y/ to the end of the /u/ such that their frequency values are acceptable for a sonorant-final /r/.

Finally, the classification of vowels as semivowels is sometimes due to intravowel energy dips. Examples of this occurrence are shown in Figure 5.17. As can be seen, an energy dip, shown in part c, occurs in the word-final /i^y/ in "guarantee" and in the second vowel of the word "explore." As a result, these portions of the vowels were analyzed by the recognition system. In the former case, the detected portion of the /i^y/ was classified as a /y/. In the latter case, the detected portion of the transcribed /ɔ/ was classified as /w-l/. Even in these instances, the classification of what may be the offglide of the /i^y/ and an inserted /w/ as semivowels is not unreasonable.

5.8 A Comparison with Previous Work

The approach and performance with respect to the recognition of semivowels of two acoustic-phonetic front ends are discussed in this section. In particular, the acoustic-

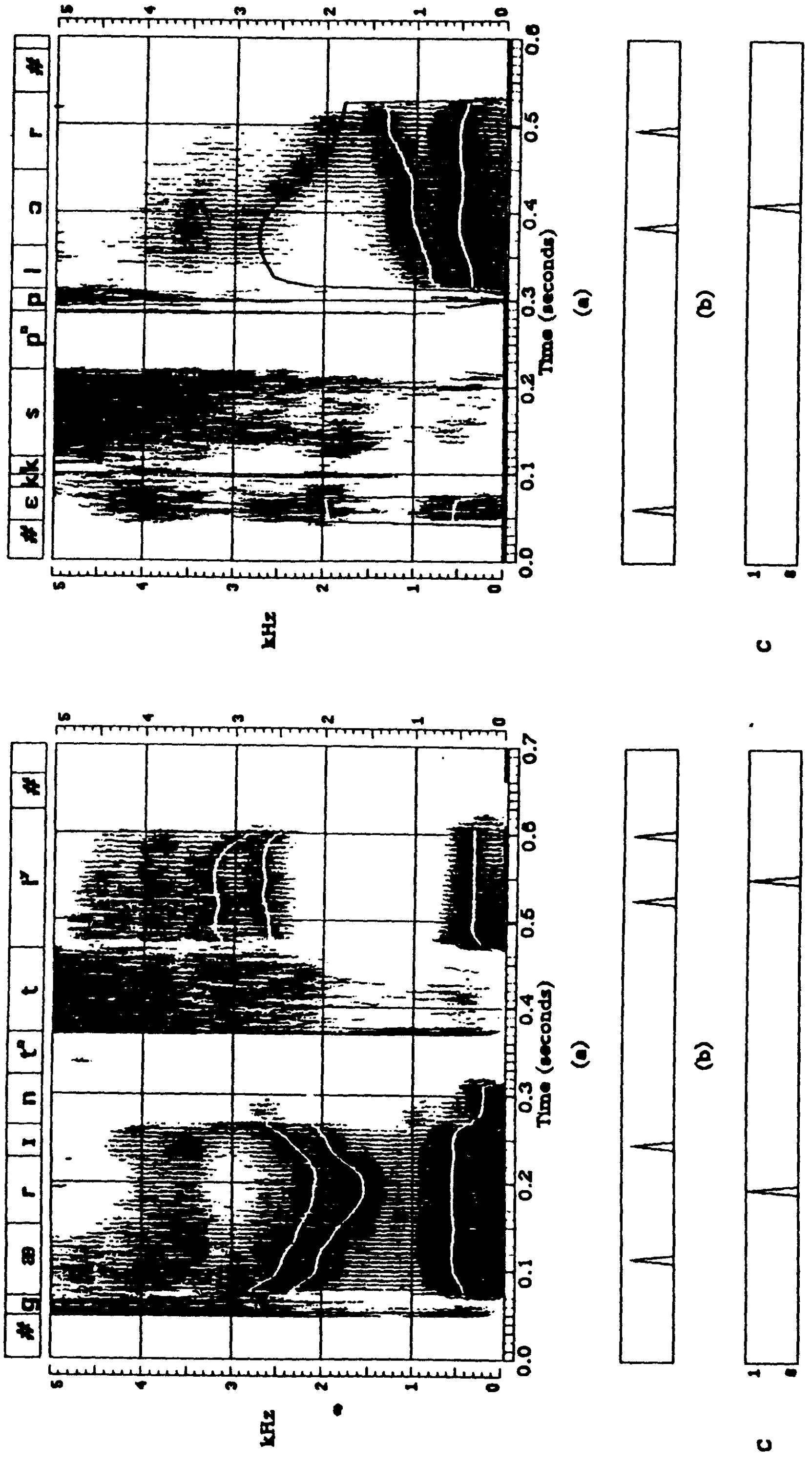


Figure 5.17: An illustration of the words "guarantee" and "explore" which contain intravowel energy dips which resulted in portions of the vowels being classified as semivowels. (a) Wide band spectrogram with formant tracks overlaid. (b) Location of energy peaks. (c) Location and confidence of energy dips.

phonetic front end developed at Lincoln Laboratories (Weinstein et al., 1975) and the acoustic-phonetic front end of the MEDRESS recognition system (Medress, 1980) are compared with the semivowel recognition system of the thesis. It is important to note that the implementation of these systems, particularly the one at Lincoln Laboratories since it was documented more thoroughly, were studied prior to initiating the present work, and in some ways guided this research.

5.8.1 LLAPFE

The semivowel recognition results obtained by LLAPFE (the Lincoln Laboratories Acoustic-Phonetic Front End) across 111 sentences spoken by six males and one female are summarized in Table 5.9. Like the data in the thesis, the results are divided on the basis of where the semivowels occurred within the voiced sonorant region. Further teasing of the data is not possible from the tabulated results.

As can be seen, LLAPFE does not attempt to recognize all semivowels occurring in all possible contexts. Although the data base contained the prevocalic /y/ in "compute," it was recognized in conjunction with the adjacent vowel. Thus, no recognition results are given for this semivowel. In addition, no attempt was made to recognize sonorant-final /r/'s. The authors felt that recognition of this sound was considerably more difficult than sonorant-initial /r/'s, since speakers will slur and sometimes omit it. Finally, the semivowels /w/ and /l/ are recognized as a single class. No further acoustic analysis is done to differentiate between them.

There are many similarities and many differences between the approaches used in LLAPFE and those used in our system. First, in both systems, the utterance is divided into sonorant and nonsonorant regions. Second, whereas recognition is divided into detection and classification in our system, these two steps are sometimes combined in LLAPFE. For intersonorant semivowels these steps are separated, whereas, for sonorant-initial and sonorant-final semivowels, they are combined. In the latter case, an /r/ identifier simultaneously segments and labels sonorant-initial /r/'s and a /w-l/ identifier simultaneously segments and labels sonorant-initial and sonorant-final /w/'s and /l/'s.

Third, both systems look for certain acoustic events to occur within semivowels. However, compared to the detection process in our system, the types of events marked in LLAPFE are not as exhaustive nor as uniform across context. For example, intersonorant semivowels are detected solely on the basis of significant energy dips (note

Table 5.9: Semivowel Recognition Results for LLAPFE: A "–" in the tables mean that the desired number could not be computed from the stated results. (Weinstein et al., 1975)

Sonorant-Initial Semivowels

	w-l	r
# tokens	–	88
undetected(%)	30	17
w-l (%)	70	0
r (%)	0	64
ɝ (%)	0	19

Intersonorant Sonorant Consonants

(*computed from the percentage of those detected)

	w-l	r	w-l + r	nasals	v,ð
# tokens	≥59	≥22	87	≥117	≥38
undetected(%)	–	–	7	–	–
w-l (%)	83*	0	56*	2*	5*
r (%)	0	91*	23*	0	0
nc (%)	17*	9*	14*	98*	95*

Sonorant-Final Semivowels

	w-l
# tokens	–
undetected(%)	30
w-l (%)	70

that intersonorant consonant clusters such as the /n/ and /l/ in "only" are treated as a single dip region). As the results show, only 93% of the intersonorant semivowels are detected in this way. This result is consistent with the data of Section 3.2.4 which show that some intervocalic semivowels which follow stressed vowels and precede unstressed vowels do not contain energy dips. The example cited by Weinstein et al. of an intersonorant semivowel which did not contain an energy dip occurs in this context. The example given is the /l/ in "millisecond." Whereas Weinstein et al. attribute the failure to detect these intervocalic semivowels to their energy dip detector, we would attribute it to the way these sounds are produced.

These detection results highlight the importance of using additional acoustic events which are based on other spectral changes. As can be seen from a comparison of the the intervocalic energy dip results of Tables 5.5 and the intersonorant energy dip results in Figure 5.9 (we are assuming that all of the intersonorant semivowels are the second member of the cluster), the detection data obtained by our system and LLAPFE are comparable. However, by combining acoustic events based on energy measures with those based on formant tracks, our system detects all of the intervocalic /w/'s, /l/'s and /r/'s occurring in all three data bases. In addition, we have found these formant minima and maxima to be particularly important in the detection of postvocalic /r/'s and /l/'s which are in clusters with other sonorant consonants. As the data of Section 3.2.6 show, these liquids do not usually contain an energy dip.

This latter point brings up another major difference between the two systems. Several cues are used in our system to detect the occurrence of more than one sound within an intersonorant dip region. However, LLAPFE treats intersonorant clusters as a single dip region. This inability to resolve both sounds in such clusters probably accounts for most of the intersonorant /w/'s and /l/'s which are misclassified as nasals. Although these confusions are not shown in Table 5.9 (they are a part of the data for "nc"), Weinstein et al. state that 12% of the intersonorant /w/'s and /l/'s were classified as nasals.

For both systems, the degree of formant movement is important for the identification of sonorant-initial and sonorant-final semivowels. Both systems look for an F3 minimum to occur within a sonorant-initial /r/. Similarly, they look for an F2 minimum to occur within a sonorant-initial and sonorant-final /w/ and /l/. However, in addition, our system looks for F3 peaks to occur within most /l/'s and within some /w/'s which are in a retroflexed environment. As can be seen in our detection data, the marking of F3 peaks is important for the detection of /l/'s. This additional acoustic cue may account for the improved recognition performance of these sounds by our system.

As in our system, LLAPFE classified the beginnings of many back vowels which are preceded by labial consonants as /w-l/. In fact, Weinstein et al. state that 27% of the sounds classified as /w-l/ were vowels preceded by /f/, /v/, /p/, /b/ or /m/.

Fourth, temporal information regarding the rate of spectral change is one of the properties used in our system to distinguish semivowels from other sounds and to distinguish between /w/'s and prevocalic /l/ allophones. Based on the classification results given in Table 5.2, this cue is useful in distinguishing between these sounds. While the time of spectral measures similar to the onsets and offsets are used in LLAPFE to segment semivowels, the values of these parameters are not used to distinguish between /w/'s and /l/'s.

Fifth, the acoustic properties in our system are directly related to specified features. Although similar measures are used in LLAPFE, no association with features is explicitly stated. In addition, the properties in our system are all based on relative measures which tend to make them speaker-independent. However, the acoustic cues used in LLAPFE are sometimes based on relative measures and sometimes based on absolute measures. Consequently, speaker dependent thresholds as well as thresholds based on the sex of the speaker are sometimes needed.

Finally, in our system, the acoustic properties are quantified using fuzzy logic such that the result is a confidence measure. Therefore, acoustic properties with different units are normalized so that they can be integrated, and the result will be another confidence measure. In addition, with this formalism, primary and secondary cues can be distinguished and qualitative descriptors can be assigned to the acoustic properties so that the rules can be easily understood. These features are not present in LLAPFE. In that system, rules are a composite of measurements and there is no convention for quantifying, on the same scale, measures with different units. Thus, /r/ and /w-l/ rules use only formant frequencies such that the application of them results in another frequency measure which does not relate directly to an acoustic event. For example, the /r/ rule in LLAPFE segments and labels a sonorant-initial /r/ if the result of the composite measurement is less than 400 Hz.

Table 5.10: Semivowel Recognition Results of the MEDRESS System: A "–" in the tables mean that the desired number could not be computed from the stated results (Medress, 1980).

	w	l	r	y
# tokens	90	164	359	37
undetected(%)	28	38	9	43
w (%)	56	–	–	–
l (%)	–	50	–	–
r(%)	–	–	85	–
y (%)	–	–	–	30

5.8.2 MEDRESS Recognition System

The semivowel recognition results obtained by the phonetic analysis component of the MEDRESS system are summarized in Table 5.10. The results given are based on the same 220 alphanumeric sequences (two, three and four words long) and data management commands used to develop the system. The utterances were spoken by three males.

Unfortunately, the semivowel recognition results are not separated on the basis of context. Furthermore, confusions between the semivowels and misclassifications of other sounds as semivowels are not given. Thus, a thorough comparison of the recognition results obtained by that system and those obtained by our system is difficult, especially in the case of /l/ and /w/. However, as can be seen from a comparison of the overall recognition results obtained by each of the data bases used in the thesis and the overall recognition results given in Table 5.10, our system does significantly better in the recognition of /r/'s and /y/'s.

In the paper describing the MEDRESS system, the discussion regarding the phonetic analysis component is brief. Therefore, an in-depth comparison of the recognition approach used in that system and that used in our system cannot be made. However, some similarities are evident. Like our system and LLAPFE, the MEDRESS system divides the speech signal into sonorant and nonsonorant regions and uses an energy dip detector to locate intersonorant semivowels. Furthermore, similar formant frequencies and movements are used to recognize the semivowels. It is not clear if minima and maxima in formant tracks are also used to detect semivowels, and it is not clear if detection and classification are done separately or simultaneously. Unlike LLAPFE and like our system, temporal information is also used to recognize /l/'s. In the MEDRESS system, this information consists of a measure which captures discontinuities in F1 at the junctures between semivowels and adjacent sounds. Finally, no speaker-dependent or sex-dependent adjustments are made.

Chapter 6

Summary and Discussion

6.1 Summary

In this thesis, we have developed a general framework for an acoustic-phonetic approach to speech recognition. This approach to recognition is based on two key assumptions. First, it assumes that phonetic segments are represented as bundles of features. Second, it assumes that the abstract features have acoustic correlates which, due to contextual influences, have varying degrees of strength. These assumptions are the basis for the framework which includes the specification of features and the determination, extraction and integration of their acoustic correlates or properties for recognition.

Although the implementation of this framework or control strategy has been tailored to the recognition of semivowels, it is based upon the general idea that the acoustic manifestation of a change in the value of a feature or group of features is marked by specific events in the sound. These acoustic events correspond to maxima or minima in particular acoustic parameters.

Thus, a major part of the control strategy of the semivowel recognition process has been to mark those acoustic events which may signal the occurrence of a semivowel. Once marked, the acoustic events are used in two ways. The time of their occurrence in conjunction with their relative strengths are used first to determine a small region from which all of the values of the acoustic properties are extracted and, second, to reduce the number of possible classifications of the detected sound. It is important to note that almost all of the acoustic properties are based on relative measures. Therefore, they tend to be independent of speaker, speaking rate and speaking level.

Although there is room for improvement in the implementation of each step in the framework, the recognition results show that the acoustic-phonetic framework is a viable methodology for speaker-independent continuous speech recognition. Fairly consistent overall recognition results in the range of 78.5% to 95% (obtained across contexts for a class consisting of both /w/ and /l/) were obtained. These results are for corpora which include polysyllabic words and sentences which were spoken by many speakers (both males and females) of several dialects. Thus, the recognition data show that much of the across-speaker variability is overcome by using a feature-based approach to recognition where relative measures are used to extract the acoustic properties.

On the other hand, there is still variability due to phenomena such as feature assimilation. In essence, the correct classification results and the misclassifications which occur show that the system is identifying patterns of features which normally correspond to semivowels. That is, many mislabelings of vowels or portions thereof and of other consonants as semivowels are caused by contextual influences and feature spreading effects which introduce feature patterns that are similar to those expected of the semivowels. These sorts of misclassifications bring into question the assignment of phonetic labels to the patterns of features. This issue is discussed in the following section.

6.2 Discussion

Throughout the thesis we have seen a number of instances of feature spreading. For example, the data of Section 3.2.4 and the recognition results given in Table 5.8 show that consonants that are normally classified as nonsonorant and voiced will sometimes appear as sonorants when they occur between vowels and/or semivowels. In addition, the feature *retroflex* appears to be highly susceptible to spreading. In this case, this phenomenon can not only result in spreading of the feature *retroflex* from an /r/ or /ɚ/ to nearby vowels and consonants, but, in certain circumstances (see Section 3.3), an underlying vowel and following /r/ can merge to form an r-colored vowel. Although it is not as clear, this same sort of phenomenon appears to occur between vowels and postvocalic, but not word-final, /l/'s as well.

Except when mergers occur, we have considered it to be an error when, due to feature spreading effects, segments that are transcribed as vowels or portions thereof, or as consonants other than semivowels, are identified by the system as semivowels. However, it is clear that in most of these cases, the sounds do have patterns of features expected for the semivowels. In fact, as was shown in Chapter 5, many segments that would be classified as semivowels in the underlying lexical representation were not transcribed as such, although they were detected and correctly classified by the system. The reasons for their exclusion from the transcription are two-fold. First, the transcription of the utterances was done in the early stages of the thesis when we did not understand as well as we do now the more subtle cues which signal the presence of a semivowel. For example, when a postvocalic /l/ follows a vowel which has many of the same properties, such as the /l/ in "wolfram," the distinguishing cue for the /l/ is often a rising third formant. Without the automatically extracted formant tracks, this F3 transition was not always apparent. Second, when we listened to the utterances, a clear semivowel is not always heard. That is, in words like "wolfram," judgement regarding the presence of an /l/ is often ambiguous. Thus, since the system sometimes recognizes semivowels which were not transcribed, but are in the underlying transcription of the utterance, it appears as if it is often correct rather than performing a misclassification, and it is probable that the transcription is incorrect instead.

Along these same lines, an analysis of some of the misclassifications of vowels as semivowels revealed that contextual influences can also result in vowel onglides and offglides which have patterns of features that normally correspond to a semivowel. That is, in the case of vowels which already have some of the features of a semivowel, adjacent sounds can cause formant movements which make portions of them look like a semivowel. These effects are apparent from many of the misclassifications listed in Appendix B. For example, across all of the speakers in Database-1 and Database-2, there are many instances where part of the transition between vowel sequences such as the transition between the /e^{y}/ and /ɪ/ in "Ghanaian," and the transition between the /$\mathrm{ɑ}^{w}$/ and /ɝ/ in "flour," were recognized as a /y/ and /w/, respectively, but were not transcribed as such. Similarly, as was shown in Chapter 5, there are several instances where sonorant-initial back vowels preceded by labial consonants are called either /w/, /l/ or /w-l/, and sonorant-initial front vowels preceded by coronal consonants are called /y/.

It is not clear that the labeling of the offglides of diphthongs as semivowels should be called an "error." In addition, it is not always clear that the labeling of the onglide of vowels as semivowels is an error. A case in point is the example shown in Figure 5.15

where the beginning of the /æ/ in "guarantee" is called a /y/. The initial segment has a high front tongue body position, leading to formant trajectories similar to those for a /y/. However, in other cases, the classification of a vowel onglide as a semivowel is not as acceptable. An example is also shown in Figure 5.15. In this case, the beginning of the first /ɔ/ in "forewarn" was labelled as a /w/. While this onglide has several acoustic properties in common with a /w/, this mislabeling is not as palatable, since /f/ and /w/ do not form an acceptable English cluster.

What these sorts of misclassifications show is that the system is recognizing certain patterns of features. In most instances, the patterns of features do correspond to a semivowel, even though some semivowels are not transcribed. However, in some instances, they do not, and it is this type of mislabeling which suggests that either labels should not be assigned to the patterns of features, or that contextual effects need to be accounted for before labeling is done.

If phonetic labels are assigned to the patterns of features, it is clear that some mechanism which accounts for feature spreading effects is needed. That is, we need to understand feature assimilation in terms of what features are prone to spreading, and in terms of the domains over which spreading occurs. In addition, techniques for dealing with other contextual influences such as those seen in the words "forewarn" and "guarantee" are needed. Such a mechanism may consist of rules which, if based on phonotactic constraints, will "clean up" phone sequences such as /fwɔ.../ so that they will appear as /fɔ.../.

If, instead of phonetic labels, lexical items are represented as matrices of features, it may be possible to avoid misclassifications due to contextual influences and feature spreading, since individual sounds are not labeled prior to lexical access. For example, consider the comparison given in Table 6.1 of what may be a partial feature matrix in the lexicon for an /ɑ/ and postvocalic /r/, with property matrices for these segments in the two repetitions of "carwash" which are shown in Figure 3.42. The lexical representation is in terms of binary features, whereas the acoustic realizations are in terms of properties whose strengths, as determined by fuzzy logic, lie between 0 and 1. We have not researched any metrics for comparing binary features and quantified properties. However, this is an important problem which needs to be solved. Instead, we will assume a simple mapping strategy where property values less than 0.5 correspond to a "−" and property values greater than or equal to 0.5 correspond to a "+."

Table 6.1: Lexical Representation vs. Acoustic Realizations of /ɑr/.

	lexical representation		realization #1		realization #2
	ɑ	r	ɑ	r	ɑ^{r}
high	−	−	0	0	0
low	+	−	1	0	1
back	+	±	1	1	1
retroflex	−	+	0	1	1

With this simple metric, a match between acoustic realization #1 and the lexical representation is straightforward. However, the mapping between acoustic realization #2 and the lexical representation is not as obvious. It may be possible for a metric to compare the two representations directly, since the primary cues needed to recognize the /ɑ/ and /r/ are unchanged. That is, the features *low* and *back* are indicative of the vowel /ɑ/ and the feature *retroflex* is indicative of an /r/ or /ɚ/. On the other hand, we may need to apply feature spreading rules before using a metric. The rules can either generate all possible acoustic manifestations from the lexical representation or generate the "unspread" lexical representation from the acoustic realization. For example, the data presented in Section 3.3 show that many r-colored vowels may underlyingly be represented by a vowel followed by /r/. Thus, acoustic realization #2 can be translated into acoustic realization #1.

In summary, many interrelated issues are highlighted by the thesis. These issues include the proper structure of the lexicon, feature assimilation, the mapping between binary features and quantified acoustic properties, and the determination, extraction and integration of the acoustic correlates of features. A fuller understanding of these matters is clearly important for an acoustic-phonetic approach to recognition and, therefore, in our opinion, they are important for speaker-independent continuous speech recognition.

6.3 Future Work

There are many directions in which this research can be extended. The issues discussed in the previous section and the analysis of the the misclassifications and no classifications in the recognition data suggest several logical extensions. In this section, we discuss some ideas and propose some experiments.

Some of the results presented in Chapter 5 show that we need a better understanding of how some features are manifested in the acoustic signal. The acoustic properties for some features are well established. However, the proper acoustic properties for others are not as clearly defined. For example, we defined the acoustic correlate of the feature "sonorant" in terms of a ratio of low frequency energy (computed from 100 Hz to 300 Hz) and high-frequency energy (computed form 3700 Hz to 7000 Hz). While the use of a parameter based on this acoustic definition resulted in the inclusion of most sonorant sounds in the detected sonorant regions, some sonorant sounds in Database-3 which had considerable high frequency energy were excluded, and a few stops with low-frequency bursts and little high frequency energy were included. Given these results, and based on our understanding of the mechanism of production of sonorant sounds, a more appropriate definition of this feature should probably be in terms of very low frequency energy. That is, it appears as if a relative measure based on only the signal energy in some range below F1 may produce better results. Clearly, much work needs to be done in determining the proper acoustic properties of some features. Knowledge gained in the areas of articulatory and perceptual correlates of features can guide this research.

The recognition data also show that some of the parameters used to capture the acoustic properties need to be refined. In some cases, there is a straightforward translation of the definition of an acoustic correlate into an adequate parameter for its extraction. However, in other cases, the transformation of an acoustic property into a reliable parameter is not as clear. Such dilemmas will probably be resolved as we gain more knowledge in areas such as auditory processing. For example, consider the formant tracker developed in the thesis. As in past attempts at formant tracking, incorrect tracks due to effects such as peak mergers, increased formant bandwidths, and nasalization are sometimes produced. The solution to this problem may be the development of a better formant tracker, or other techniques which extract the same sort of spectral information (e.g., Seneff (1987) has developed an auditory-based technique which extracts "line-formants," straight-line segments which sketch out the formant trajectories without explicitly labelling F1, F2, F3, etc.). On the other hand, the solution to this problem may be the use of additional measures, such as spectral tilt and the frequency range of the major spectral prominence, in conjunction with formant tracks. Such measures do not require the resolution of spectral peaks. Thus, their use in regions where formant tracks are likely to be incorrect (e.g., in nonsyllabic regions where, due to a constriction, formants may come together or their bandwidths may increase) may give better results.

A better understanding of the acoustic properties for features and parameters from which they can be reliably extracted will not only improve the performance of the present recognition system, but will also allow for the natural extension of this approach to the recognition of other sounds, including the devoiced and nonsonorant semivowel allophones. The addition of other features should also reduce the misclassifications of other consonants as semivowels.

Another extension along the same line is an investigation of the confusions made between semivowels. The recognition data show that in some contexts, there is considerable confusion between /w/ and /l/, and, to a smaller extent, between /w/ and /r/. Perceptual tests where different acoustic cues can be manipulated and further acoustic analysis of the sounds which were confused may reveal additional or more appropriate acoustic cues needed to make these distinctions. In addition, such research may give insights into how the different acoustic properties should be integrated. That is, such a study may allow for the distinction between acoustic properties which are primary and those which are secondary.

Finally, feature assimilation and lexical representation are important issues which need to be better understood. The mapping between the acoustic signal which contains the effects of spreading phenomenon and items in the lexicon is a difficult and important problem. The recognition results of Chapter 5 lead us to believe that the proper representation of lexical items is in terms of feature matrices. Thus, we need to develop techniques for accessing lexical items, which are represented by binary features, from quantified acoustic properties which, due to phenomena such as feature spreading, have varying degrees of strength and extent over time. Spectrogram reading provides an expedient framework in which this question can be studied, since it eliminates the problem of computer extraction of the acoustic properties. That is, the acoustic properties can be identified from this visual representation. In attempting to compare the lexical items and the extracted acoustic properties, several issues will have to be addressed. First, the proper structure of these representations must be developed. For example, whereas lexical items are represented in terms of matrices of features, it is probable that some further structure is imposed on these matrices, taking into account what is known about syllable structure, larger units such as words and feet, relations between features, etc. Certainly, units larger than segments are needed to adequately capture contextual influence. Thus, for example, the feature matrix may consist of columns which describe the transitions between the phonetic segments as well as a column for the steady state characteristics of the sounds. Second, this type of spectrogram reading experiment should give some insight into the features which are susceptible to spreading and the contexts in which spreading is likely to occur. Finally, this experiment should help to determine whether or not feature spreading rules are needed or if this phenomenon can be accounted for in a natural way without elaborate rules.

REFERENCES

Barnett, J.A., Bernstein, M.I., Gillman, R.A. and Kameny, I.M.,"The SDC Speech Understanding System," *Trends in Speech Recognition*, ed. Wayne A. Lea, New Jersey: Prentice-Hall, Inc., 1980.

Bickley, C. and Stevens, K.N., "Effects of a vocal tract constriction on the glottal source: data from voice consonants," ***Laryngeal Function in Phonation and Respiration*** eds: T. Baer, C. Sasaki and K. Harris, San Diego: College Hill Press, pp. 239-253, 1987.

Bladon, R.A.W. and Al-Bamerni, Ameen, "Coarticulation Resistance in English /l/," *Journal of Phonetics*, vol. 4, pp. 137-150, 1976.

Bond, Z.S., "Identification of Vowels Excerpted from /l/ and /r/ Contexts," *J. Acoust. Soc. Am.*, vol. 60, pp. 906-910, October 1976.

Chomsky, N. and Halle, M. *The Sound Pattern of English*, New York: Harper and Row, 1968.

Christensen, R.L., Strong, W.J., and Palmer, E.P., "A Comparison of Three Methods of Extracting Resonance Information from Predictor-Coefficient Coded Speech," *IEEE Trans. Acoust., Speech, and Sig. Proc.*, vol. 24, pp. 8-14, February 1976.

Cole, R.A., Stern, R.M., Phillips, M.S., Brill, S.M., Pilant, A.P. and Specker, P., "Feature-based speaker-independent recognition of isolated letters," *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, vol 2., pp. 731-733, 1983.

Coler, C.R., Huff, E.M., Plummer, R.P., and Hitchcock, M.H., "Automatic Speech Recognition Research at NASA-Ames Research Center," *Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application*, ed. R. Breaux, M.Curran, and E. Huff, NASA Ames Research Center, Moffett Field, Ca, pp. 171-196.

Cutler, A. and Foss, D., "On the Role of Sentence Stress in Sentence Processing,". *Language and Speech*, vol. 20, pp. 1-10, 1977.

Dalston, R.M., "Acoustic Characteristics of English /w,r,l/ Spoken Correctly by Young Children and Adults," *J. Acoust. Soc. Am.*, vol. 57 no. 2, pp. 462-469, February 1975.

Davis, K., Biddulph,R.., and Balashek,S., "Automatic Recognition of Spoken Digits, *J. Acoust. Soc. Am.*, vol. 24 no. 6, pp. 637-642, November, 1952.

De Mori, Renato, *Computer Models of Speech Using Fuzzy Algorithms.* New York and London: Plenum Press, 1983.

Doddington, G.R., "Personal Identity Verification Using Voice," presented at ELECTRO 76, Boston, Mass., 1976.

Erman, L.D. and Lesser, V.R., "The HEARSAY-II Speech Understanding System: A Tutorial," *Trends in Speech Recognition*, ed. Wayne A. Lea, New Jersey: Prentice-Hall, Inc., 1980.

Fant, Gunnar, *Acoustic Theory of Speech Production.* The Netherlands: Mouton & Co., 1960.

Gold, B. and Rabiner, L., "Parallel Processing Techniques for Estimating Pith Periods of Speech in the Time Domain," *J. Acoust. Soc. Am.*, vol. 46, no. 2, pp. 442-449, 1969.

Huttenlocher, D. and Zue, V., "Phonotactic and Lexical Constraints in Speech Recognition," Proc. of the National Conference on Artificial Intelligence, pp. 172-176, August, 1983.

Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. 23, pp 67-72, 1975.

Jakobson, R., Fant, G. and Halle, M., "Preliminaries to Speech Analysis," *MIT Acoustics Lab. Tech. Rep. No. 13*, 1952.

Jelinek, F., Bahl, L.R., and Mercer, R.L., "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech," *IEEE Trans. on Infor. Theory*, vol. IT-21, pp. 250-256, 1975.

Jelinek, F., "Continuous Speech Recognition by Statistical Methods," *Proceedings of the IEEE*, vol. 64, no. 5, pp. 532-556, 1976.

Jelinek, F., "Self-Organized Continuous Speech Recognition," *Proceedings of the NATO Advanced Summer Inst. Auto. Speech Analysis and Recognition*, France, 1981.

Kameny, I., "Automatic Acoustic-Phonetic Analysis of Vowels and Sonorants," *IEEE Internat. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 166-169, 1976.

Kameny, I., "Comparison of the Formant Spaces of Retroflexed and Non-retroflexed Vowels," *IEEE Symp. Speech Recog.*, pp. 80-T3 - 84-T3, 1974.

Klatt, D.H., "Review of the ARPA Speech Understanding Project," *J. Acoust. Soc. Am.*, vol. 62, no. 6, pp. 1345-1366, December 1977.

Lamel, L., Kassel, R., and Seneff, S., "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. Speech Recog. Workshop*, CA., 1986.

Lea, W.A., "Speech Recognition: What is Needed Now?," *Trends in Speech Recognition*, ed. Wayne A. Lea, New Jersey: Prentice-Hall, Inc., 1980.

Lea, W.A., "Speech Recognition: Past, Present and Future," *Trends in Speech Recognition*, ed. Wayne A. Lea, New Jersey: Prentice-Hall, Inc., 1980.

Lehiste, I., "Acoustical Characteristics of Selected English Consonants," *Report No. 9*, University of Michigan, Communication Sciences Laboratory, Ann Arbor, Michigan, July 1962.

Lehiste, I. and Peterson, G.E., "Transitions, Glides, and Diphthongs," *J. Acoust. Soc. Am.*, vol. 33, no. 3, pp. 268-277, March 1961.

Leung, H. and Zue, V.W., "A Procedure for Automatic Alignment of Phonetic Transcription with Continuous Speech," *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 2.7.1-2.7.4, 1984.

Lindgren, N., "Machine Recognition of Human Language, Part I," *IEEE Spectrum*, vol. 2, pp. 114-136, 1965.

Lisker, L., "Minimal Cues for Separating /w,j,r,l/ in Intervocalic Position," *Word*, vol. 13, pp. 256-267, 1957.

Lowerre, B., "The Harpy Speech Recognition Systems," Ph.D. dissertation, Computer Science Dept., Carnegie-Mellon U., 1977.

Martin, T.B., Nelson, A.L. and Zadell, H.J., "Speech Recognition by Feature Abstraction Techniques, *Technical Report No. AL TDR 64-176*, RCA, Camden, New Jersey, 1964.

Martin, T.B., "Practical Applications of Voice Input to Machines," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 487-501.

McCandless, Stephanie, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. 22, no. 2, pp 135-141.

McGovern, Katherine and Strange, Winfred, "The Perception of /r/ and /l/ in Syllable-initial and Syllable-final Position," *Perception and Psychophysics*, vol. 21 no. 2, pp. 162-170, 1977.

Medress, M.F., "The Sperry Univac System for Continuous Speech Recognition," *Trends in Speech Recognition*, ed. Wayne A. Lea, New Jersey: Prentice-Hall, Inc., 1980.

Medress, M.F., "Computer Recognition of Single-Syllable English Words," Ph.D. Thesis, Massachusetts Institute of Technology, 1969.

Mermelstein, P., "Automatic Segmentation of Speech into Syllabic Units," *J. of Acoust. Soc. Am.*, vol. 58, pp. 880-883, 1975.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. Jenkins, J. and Fujimura, O., "An effect of Linguistic Experience: The Discrimination of [r] and [l] by Native Speakers of Japanese and English," *Perception and Psychophysics*, vol. 18, no. 5, pp. 331-340, 1975.

Mochizuki, M., "The Identification of /r/ and /l/ in Natural and Synthesized Speech," *Journal of Phonetics*, vol. 9, pp. 283-303, 1981.

Myers, C.S. and Rabiner, L.R., "A Level Building Dynamic Time Warping Algorithm for Connected Word-Recognition," *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 951-955, 1981.

Nakatani, L.H. and Dukes, K.D., "Locus of Segmental Cues for Word Juncture," *J. Acoust. Soc. Am.*, vol. 62, no. 3, pp. 714-719, September 1977.

O'Connor, J.D., Gertsman, L.J., Liberman, A.M., Delattre, P.C., and Cooper, F.S., "Acoustic Cues for the Perception of Initial /w,j,r,l/ in English, *Word*, vol. 13, pp. 24-43, 1957.

Prazdny, K., "Waveform Segmentation and Description Using Edge Preserving Smoothing," *Computer Vision, Graphics, and Image Processing*, vol 23, pp. 327-333, YEAR?

Sakoe, H. and Chiba, S., "A Dynamic Programming Approach to Continuous Speech Recognition," *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 65-68, 1971.

Rabiner, L.R., Levinson, S.E., Rosenberg, A.E., Wilpon, J.G., "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques," *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. 27, no. 4, pp. 336-349, 1979.

Rabiner, L.R., Levinson, S.E., and Sondhi, M.M., "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition," *Bell System Technical Journal*, vol. 62, no. 4, 1983.

Schwartz, R., Chow, Y., Kimball, O., Roucous, S., Krasner, M. and Makhoul, J., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, vol 3., pp. 1205-1208, 1985.

Schwartz, R. and Makhoul, J., "Where the Phonemes Are: Dealing with Ambiguity in Acoustic-Phonetic Recognition," *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. 23, no. 1, pp. 50-53, February 1975.

Selkirk, E.O., "The Syllable," *The Structure of Phonological Representations (part II)*, ed. H. van der Hulst and N. Smith, Dordrecht: Foris Publications, 1982.

Seneff, Stephanie, "A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research," *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 1983-1986, 1986.

Seneff, Stephanie, "Vowel Recognition based on Line-Formants derived from an Auditory-Based Spectral Representation," to be presented at the *Eleventh International Congress of Phonetic Sciences*, Estonia, USSR, August 1987.

Shafer, Ronald and Rabiner, Lawrence, "System for Automatic Formant Analysis of Voiced Speech," *J. Acoust. Soc. Am.*, vol. 47, no. 2, July 1969, pp. 634-648.

Stevens, K.N., "Models of Phonetic Recognition II: An Approach to Feature-Based Recognition," Proc. of the Montreal Symp. on Speech Recog., pp. 67-68, July 1986.

Stevens, K.N., Keyser, S.J. and Kawasaki, H., "Toward a Phonetic and Phonological Theory of Redundant Features," *Invariance and Variability in Speech Processes*, eds. J.S. Perkell and D.H. Klatt, New Jersey: Lawrence Erlbaum Associates, pp. 426-449, 1986.

Stevens, K.N., book on acoustic phonetics to be published.

Trager, G.L. and Smith, H.L. Jr., "An Outline of English Structure," *Studies in Linguistics: Occasional Papers 3*, Norman, Oklahoma: Battenburg Press, 1951.

Weinstein, C.J., McCandless, S.S., Mondshein, L.F. and Zue V.W., "A System for Acoustic Phonetic Analysis of Continuous Speech," *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. 23, no. 1, pp. 54-67, February 1975.

Wiren, J. and Stubbs, H., "Electronic Binary Selection System for Phoneme Classification," *J. Acoust. Soc. Am.*, vol. 28, pp. 1082-1091, 1956.

Woods, W., Bates, M., Brown, G., Bruce, B., Cook, C., Klovstad, J., Makhoul, J., Nash-Webber, B., Schwartz, R., Wolf, J. and Zue, V., "Speech Understanding Systems," *BBN Report No. 3438*, Bolt Beranek and Newman Inc., Cambridge, MA., December 1976.

Yegnanarayana, B., "Formant Extraction from Linear-Prediction Phase Spectra," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1638-1640, May 1978.

Zue, V.W. and Cole, R.A.,"Experiments on Spectrogram Reading," *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 116-119, April 1979.

Zue, V.W., "Speech Spectrogram Reading: An acoustic Study of English Words and Sentences," *MIT Special Summer Course*, Cambridge, Ma., 1985.

Zue, V., Cyphers, D., Kassel, R., Kaufman, D. Leung, H., Randolph, M. Seneff, S., Unverferth, J., and Wilson, T. "The Development of the MIT LISP-Machine Based Speech Research Workstation," *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 7.6.1-7.6.4, 1986.

Appendix A

Corpus of Polysyllabic Words

Table A.1: Alphabetical listing of the polysllabic words and their phonemic transcriptions. The transcriptions, originally from the Merriam-Webster Pocket Dictionary, were checked to ensure consistency.

Words	**Phonemic Transcription**
African	’æfrɪkɨn
afterward	’æftɚwɚd
airline	’ær∗l‘ɑʸn
albatross	’ælbətr‘ɔs
almost	’ɔlm‘oʷst
already	ɔlr’ɛdiʸ
always	’ɔlwəz
anthrax	’ænθr‘æks
Aquarius	əkw’æriʸəs
arteriosclerosis	ɑrt‘ɪriʸoʷsklər’oʷsɨs
assuage	əsw’eʸǰ
astrology	əstr’ɑləǰiʸ
bailiwick	b’eʸlɪw‘ɪk
banyan	b’ænyən
beauty	by’utiʸ
behavior	bɪh’eʸvyɚ
bellwether	b’ɛlw‘ɛðɚ
bewail	biʸw’eʸl
bless	bl’ɛs
blurt	bl’ɝt
bourgeois	b‘ʊržw’ɑ
brilliant	br’ɪlyənt
bucolic	byuk’ɑlɪk
bulrush	b’ʊlr‘ʌš

Word	**Phonemic Transcription**
bureaucracy	byɚ’ɑkrɨsi^{y}
bureaucratic	by‘ɜəkr’ætɪk
bushwhack	b’ʊš*hw‘æk
calculus	k’ælkyʊləs
caloric	kəl’ɔrɪk
canalize	kən’æl‘ɑ^{y}z
carwash	k’ɑrw‘ɑš
cartwheel	k’ɑrt*hw‘i^{y}l
cellular	s’ɛlyʊlɚ
chignon	š’i^{y}y‘ɑn
chivalric	šəv’ælrɪk
chlorination	kl‘o^{w}rən’e^{y}šɨn
choleric	k’ɑlərɪk
clean	kl’i^{y}n
clear	kl’ɪr
cognac	k’o^{w}ny‘æk
coiffure	kwɑfy’ʊr
conflagration	k‘ɑnfləgr’e^{y}šɨn
contrariwise	k’ɑntr‘ɛri^{y}w‘ɑ^{y}z
cordwainer	k’ɔrdw‘e^{y}nɚ
correlation	k‘ɔrəl’e^{y}šɨn
cream	kr’i^{y}m
cumulative	ky’umyʊlɨtɪv
curator	kyʊr’e^{y}tɚ
cutthroat	k’ʌt*θr‘o^{w}t
darwin	d’ɑrwɨn
demoralize	dəm’ɔrəl‘ɑ^{y}z
derogatory	dɪr’ɑgɨto^{w}ri^{y}
devoir	dəvw’ɑr
dillydally	d‘ɪli^{y}*d’æli^{y}
dislocate	d’ɪslo^{w}k‘e^{y}t
disqualify	d‘ɪskw’ɑləf‘ɑ^{y}
disquisition	d‘ɪskwəz’ɪšɨn

Word	**Phonemic Transcription**
disreputable	d‘ɪsr’ɛpyutəbl̩
diuretic	d‘ɑ^{y}yʊr’ɛtɪk
donnybrook	d’ɑni^y*br‘ʊk
dossier	d’ɔsy‘e^y
dramatic	drəm’ætɪk
dwell	dw’ɛl
ellwood	’ɛlwʊd
emasculate	ɪm’æskyul‘e^{y}t
ennui	’ɑnw‘i^y
enshrine	ɪnšr’ɑ^{y}n
esquire	’ɛskw‘ɑ^{y}r
Eurasian	yʊr’e^{y}žɨn
eurologist	yʊr’ɑləǰɨst
everyday	’ɛvri^y*d‘e^y
exclaim	ɪkskl’e^{y}m
exclusive	ɪkskl’usɪv
exploitation	‘ɛkspl‘ɔ^{y}t’e^{y}šɨn
explore	ɪkspl’o^{w}r
expressway	ɪkspr’ɛs*w‘e^y
exquisite	ɛkskw’ɪzɨt
extraordinarily	ɪkstr‘ɔrdn̩’ɛrəli^y
extrapolate	ɪkstr’æpəl‘e^{y}t
familiarity	fəm‘ɪly’ærəti^y
farewell	fær*w’ɛl
fibroid	f’ɑ^{y}brɔ^{y}d
flamboyant	flæmb’ɔ^{y}ənt
flirt	fl’ɝt
flour	fl’ɑ^{w}r
flourish	fl’ʊrɪš
fluorescence	flʊr’ɛsn̩s
foreswear	fɔrsw’ær
forewarn	fo^{w}rw’ɔrn
fragrant	fr’e^{y}grənt

Word	**Phonemic Transcription**
fraudulent	fr'ɔǰʊlənt
frivolous	fr'ɪvl̩əs
froward	fr'o^{w}wɚd
frustration	fr'ʌstr'e^{y}šɨn
fuel	fy'ul
Ghanaian	gɑn'e^{y}yən
gladiolus	gl'ædi^{y}'o^{w}ləs
glass	gl'æs
granular	gr'ænyʊlɚ
grizzly	gr'ɪzli^{y}
guarani	gw'ɑrən'i^{y}
guarantee	g'ærənt'i^{y}
harlequin	h'ɑrlɪkwən
harmonize	h'ɑrmən'ɑ^{y}z
heirloom	'ærl'um
heroin	h'ɛro^{w}ɨn
horology	ho^{w}r'ɑləǰi^{y}
humiliate	hyum'ɪli^{y}'e^{y}t
incredulously	'ɪnkr'ɛǰʊləsli^{y}
infrequently	'ɪnfr'i^{y}kwəntli^{y}
interweave	'ɪntɚw'i^{y}v
inward	'ɪnwɚd
Israelite	'ɪzri^{y}əl'ɑ^{y}t
kyat	ki^{y}y'ɑt
laceration	l'æsɚ'e^{y}šɨn
leapfrog	l'i^{y}p*fr'ɔg
legalistic	l'i^{y}gəl'ɪstɪk
legislation	l'ɛǰɨsl'e^{y}šɨn
librarian	lɑ^{y}br'eri^{y}ən
linguistics	liŋgw'istiks
livelihood	l'ɑ^{y}vli^{y}h'ud
loathly	l'o^{w}dli^{y}
locale	lo^{w}k'æl

Word	**Phonemic Transcription**
luxurious	l‘ʌgž’ʊri^{y}əs
mansuetude	m’ænswɪt‘ud
marijuana	m‘ærəw’ɑnə
marlin	m’ɑrlɨn
memoir	m’ɛmw‘ɑr
menstrual	m’ɛnstrʊl̩
miniscule	m’ɪnɨsky‘ul
miscue	m‘ɪsky’u
misquotation	m‘ɪskwo^{w}t’e^{y}šɨn
misquote	m‘ɪskw’o^{w}t
misrule	m‘ɪsr’ul
muscular	m’ʌskyʊlɚ
musculature	m’ʌskyʊləč‘ʊr
northward	n’ɔrθwɚd
Norwegian	nɔrw’i^{y}ǰɨn
oneself	w‘ʌn∗s’ɛlf
onslaught	’ɑn∗sl‘ɔt
ornery	’ɔrnɚi^{y}
periwig	p’ɛrɪw‘ɪg
picayune	p‘ɪki^{y}y’un
plurality	plʊr’ælɨti^{y}
poilu	pwɑl’u
pollywog	p’ɔli^{y}w‘ɑg
postlude	p’o^{w}stl‘ud
postwar	p‘o^{w}stw’ɔr
prime	pr’ɑ^{y}m
promiscuously	prəm’ɪskyʊəsli^{y}
pule	py’ul
puree	pyʊr’e^{y}
purulent	py’ʊrl̩ənt
quadruplet	kwɑdr’ʌplɨt
quarry	kw’ɔri^{y}
queen	kw’i^{y}n
queer	kw’ɪr

Word	**Phonemic Transcription**
queue	ky’u
quotation	kwo^{w}t’e^{y}šɨn
radiology	r‘e^{y}di^{y}’aləǰi^{y}
rationale	r‘æšn̩’æl
rauwolfia	ra^{w}w’ʊlfi^{y}ə
reconstruct	r‘ɪkɨnstr’ʌkt
requiem	r’ɛkwi^{y}əm
resplendent	rɪspl’ɛndɨnt
reunion	r‘i^{y}y’unyən
rhinoceros	ra^{y}n’asɚəs
ringlet	r’ɪŋlɨt
riyal	ri^{y}y’ɔl
roulette	rul’ɛt
rule	r’ul
scroll	skr’o^{w}l
seaward	s’i^{y}wɚd
shrill	šr’ɪl
silhouette	s‘ɪlo^{w}’ɛt
skew	sky’u
sling	sl’ɪŋ
slop	sl’ap
snarl	sn’arl
soliloquize	səl’ɪləkw‘a^{y}z
splenetic	splɪn’ɛtɪk
splice	spl’a^{y}s
spurious	spy’ʊri^{y}əs
squall	skw’ɔl
square	skw’ær
squeamish	skw’i^{y}mɪš
stalwart	st’ɔlwɚt
Swahili	swah’i^{y}li^{y}
swap	sw’ap
swing	sw’ɪŋ
swirl	sw’ɝl

Word	**Phonemic Transcription**
swollen	sw’o^{w}lɨn
swung	sw’ʌŋ
thwart	θw’ɔrt
transcribe	trænskr’ɑ^{y}b
twain	tw’e^{y}n
twilight	tw’ɑ^{y}l‘ɑ^{y}t
ukulele	y‘ukəl’e^{y}li^{y}
unaware	‘ʌnəw’ær
unctuous	’ʌŋkčəwəs
unilateral	y‘unəl’ætəl̩
unreality	‘ʌnri^{y}’ælɨti^{y}
urethra	yʊr’i^{y}θrə
vuvula	y’uvyʊlə
view	vy’u
volume	v’ɑlyʊm
voluntarily	v‘ɑlɨnt’ɛrəli^{y}
voyageur	vw‘ɑy‘ɑž’ɝ
wagonette	w‘ægən’ɛt
wallflower	w’ɔl*fl‘ɑ^{w}ɚ
Walloon	wɑl’ʊn
walnut	w’ɔln‘ʌt
walrus	w’ɔlrɨs
waterproof	w’ɔtɚ*pr‘uf
weatherworn	w’ɛðɚw‘o^{w}rn
whippoorwill	hw’ɪpɚw‘ɪl
whitlow	hw’ɪtl‘o^{w}
widespread	w’ɑ^{y}d*spr‘ɛd
willowy	w’ɪlo^{w}i^{y}
withdraw	w‘ɪθ*dr’ɔ
withhold	wɪθ*h’o^{w}ld
wolfram	w’ʊlfrəm
wolverine	w‘ʊlvɚ’i^{y}n
worthwhile	w‘ɝθ*hw’ɑ^{y}l
wristlet	r’ɪstlɨt

Word	Phonemic Transcription
wrought	r’ɔt
yawl	y’ɔl
yell	y’ɛl
yearlong	y’ɪr*l‘ɔŋ
yon	yɑn
yore	y’o^{w}r

Table A.2: Word-initial semivowels which are adjacent to stressed vowels.

w	l	r	y
wallflower	leapfrog	requiem	uvula
walnut	livelihood	ringlet	yearlong
walrus	loathly	wristlet	yell
waterproof		rule	yawl
weatherworn		wrought	yon
widespread			
willowy			
wolfram			
wristlet			

Table A.3: Word-initial semivowels which are adjacent to vowels which are either unstressed or have secondary stress.

w	l	r	y
Walloon	librarian	rauwolfia	Eurasian
withhold	linguistics	resplendent	eurologist
	locale	rhinoceros	urethra
		riyal	
		roulette	

Table A.4: Prevocalic semivowels that are adjacent to a fricative and adjacent to a stressed vowel.

w	l	r	y
assuage	flirt	disreputable	coiffure
devoir	flour	enshrine	fuel
foreswear	flourish	fragrant	view
swap	legislation	fraudulent	
swing	sling	frivolous	
swirl	slop	froward	
swollen		infrequently	
swung		misrule	
thwart		shrill	
whippoorwill			
whitlow			
worthwhile			
bourgeois			

Table A.5: Prevocalic semivowels that are adjacent to a fricative and adjacent to a vowel which is either unstressed or has secondary stress.

w	l	r	y
bushwhack	conflagration	anthrax	behavior
cartwheel	dislocate	cutthroat	dossier
mansuetude	flamboyant	frustration	humiliate
northward	fluorescence	leapfrog	uvula
Swahili	grizzly	African	
voyageur	incredulously	everyday	
	livelihood	Israelite	
	loathly	urethra	
	onslaught	wolfram	
	promiscuously		
	wallflower		

Table A.6: Prevocalic semivowels which are adjacent to a stop and adjacent to a stressed vowel.

w	l	r	y
Aquarius	bless	brilliant	cumulative
dwell	blurt	bureaucratic	pule
linguistics	clean	conflagration	purulent
quarry	clear	cream	queue
queen		granular	
queer		grizzly	
twain		incredulously	
twilight		librarian	
		quadruplet	
		withdraw	

Table A.7: Prevocalic semivowels which are adjacent to a stop and adjacent to a vowel which is either unstressed or has secondary stress.

w	l	r	y
coiffure	chlorination	albatross	bucolic
cordwainer	gladiolus	bureaucracy	bureaucracy
guarani	infrequently	contrariwise	bureaucratic
harlequin	plurality	donnybrook	calculus
infrequently	quadruplet	dramatic	curator
poilu	whitlow	fibroid	disreputable
quadruplet		fragrant	puree
quotation		waterproof	
requiem			
soliloquize			

Table A.8: Prevocalic semivowels which are adjacent to a fricative-stop cluster and adjacent to a stressed vowel.

w	l	r	y
disqualify	exclaim	astrology	miscue
exquisite	exclusive	expressway	spurious
misquote	explore	extrapolate	skew
postwar	resplendent	frustration	
squall	splice	reconstruct	
square		scroll	
squeamish		transcribe	

Table A.9: Prevocalic semivowels which are adjacent to a fricative-stop cluster and adjacent to a vowel which has secondary stress.

w	l	r	y
esquire	exploitation	extraordinarily	miniscule
	postlude	widespread	

Table A.10: Prevocalic semivowels which are adjacent to a fricative-stop cluster and adjacent to unstressed vowels.

w	l	r	y
disquisition	arteriosclerosis	menstrual	emasculate
misquotation	splenetic		muscular
	wristlet		musculature
			promiscuously

Table A.11: Intervocalic Semivowels which occur before stressed vowels.

w	l	r	y
bewail	caloric	arteriosclerosis	kyat
interweave	correlation	curator	picayune
marijuana	legalistic	derogatory	reunion
rauwolfia	roulette	diuretic	riyal
unaware	soliloquize	Eurasian	
	ukulele	eurologist	
	unilateral	fluorescence	
	Walloon	horology	
		plurality	
		urethra	

Table A.12: Intervocalic Semivowels which follow vowels which are stressed.

w	l	r	y
froward	astrology	Aquarius	Ghanaian
seaward	bailiwick	caloric	
	bucolic	demoralize	
	choleric	extraordinarily	
	disqualify	familiarity	
	eurologist	flourish	
	gladiolus	heroin	
	horology	librarian	
	humiliate	luxurious	
	plurality	periwig	
	pollywog	purulent	
	radiology	spurious	
	soliloquize	voluntarily	
	swollen		
	unreality		
	willowy		

Table A.13: Intervocalic Semivowels which occur between unstressed vowels.

w	l	r	y
afterward	calculus	chlorination	diuretic
unctuous	cumulàtive	choleric	
	dillydally	contrariwise	
	fraudulent	correlation	
	incredulously	guarani	
	musculature	marijuana	
	silhouette		
	voluntarily		

Table A.14: Intersonorant Semivowels which are adjacent to other semivowels.

rw	rl	lr	lw	ly
carwash	harlequin	bulrush	bellwether	brilliant
Darwin	marlin	chivalric	stalwart	cellular
forewarn	snarl	walrus	Ellwood	volume
Norwegian	airline			
	heirloom			
	yearlong			

Table A.15: Intersonorant Semivowels which are adjacent to nasals.

w	l	r	y
ennui	walnut	forewarn	banyan
inward	almost	harmonize	granular
memoir	ringlet	unreality	chignon
		weatherworn	cumulative

Table A.16: Word-final semivowels.

l	r
bewail	clear
cartwheel	coiffure
dwell	devoir
farewell	esquire
fuel	explore
locale	flour
miniscule	foreswear
misrule	memoir
pule	musculature
rationale	postwar
riyal	queer
shrill	square
squall	unaware
swirl	
whippoorwill	
worthwhile	
rule	
yawl	
yell	

Table A.17: Postvocalic semivowels which are not word-final.

l	r
oneself	bourgeois
wolfram	foreswear
wolverine	northward
withhold	cartwheel
	cordwainer
	thwart

Table A.18: Word-initial vowels.

tense	lax
African	Aquarius
afterward	assuage
airline	astrology
albatross	Ellwood
almost	emasculate
already	enshrine
always	esquire
anthrax	everyday
arteriosclerosis	exclaim
ennui	exclusive
heirloom	exploitation
onslaught	explore
ornery	expressway
	exquisite
	extraordinarily
	extrapolate
	incredulously
	infrequently
	interweave
	inward
	Israelite
	unaware
	unctuous
	unreality

Table A.19: Word-initial nasals and /h/'s.

m	n	h
mansuetude	northward	harlequin
marijuana	Norwegian	harmonize
marlin		heroin
memoir		horology
menstrual		humiliate
miniscule		
miscue		
misquotation		
misquote		
misrule		
muscular		
musculature		

Table A.20: Intervocalic nasals and /h/'s.

m	n	h
demoralize	canalize	withhold
dramatic	chlorination	behavior
familiarity	donnybrook	livelihood
humiliate	Ghanaian	Swahili
promiscuously	harmonize	
squeamish	miniscule	
	rhinoceros	
	splenetic	
	unilateral	
	wagonette	

Table A.21: Word-final nasals.

m	n	ng
cream	African	sling
exclaim	airline	swing
heirloom	banyan	swung
requiem	chignon	
volume	chlorination	
wolfram	clean	
	conflagration	
	correlation	
	Darwin	
	disquisition	
	enshrine	
	Eurasian	
	exploitation	
	frustration	
	Ghanaian	
	harlequin	
	heroin	
	laceration	
	legislation	
	librarian	
	marlin	
	misquotation	
	Norwegian	
	picayune	
	queen	
	quotation	
	reunion	
	swollen	
	twain	
	Walloon	
	wolverine	

Appendix B

Vowels Misclassified as Semivowels

The following list of words contains a sample of vowel onglides and vowel offglides which were recognized as semivowels. The portion of the vowel which was "misclassified" as a semivowel can be inferred from the phonemes within the parenthesis following the words.. These sounds surround the vowel onglide or vowel offglide. Thus, the phonemes (/bu/) after the word "bourgeois" in the column labeled "w,w-l,l" indicate that the beginning portion of the vowel /u/ was sometimes recognized as /w/, /w-l/ and /l/. Similarly, the phonemes (/i^{y}ʌ/) after the word "aquarius" in the column labeled "y" indicate that in one or more repetitions of this word, a /y/ was not transcribed, but the offglide of the /i^y/ was recognized as a /y/. Note that the symbol "#" is sometimes included in the parenthesis. This symbol denotes a word boundary. Thus, the (/ɝ#/) following the word "behavior" in the "r" column means that in one or more repetitions of this word, the last sound was transcribed as an /ɝ/, but was recognized as an /r/. Finally, in examples of "misclassifications" of vowel portions as /r/ which involve spreading of the feature retroflex, three sounds are in the parenthesis. As in the other cases, the sounds surrounding the vowel portion classified as /r/ are given. However, to mark the direction of feature spreading, the position of the /r/ or /ɝ/ with respect to these sounds is also shown.

Table B.1: Portions of vowels which were classified as a semivowel.

w,w-l,l	l	r	y
bourgeois (/bʊ/)	albatross (/ɔb/)	african (/æfr/)	aquarius (/i^{y}ʌ/)
bulrush (/bʊ/)	almost (/ɔm/)	aquarius (/ɝɪ/)	arteriosclerosis (/tɪ/)
bushwhack (/bʊ/)	always (/ɔw/)	behavior (/ɝʉ/)	arteriosclerosis (/i^{y}o^{w}/)
foreswear (/fɔ/)	disqualify (/fɑ^{y}/)	cellular (/ɝʉ/)	astrology (/ǰi^{y}/)
forewarn (/fɔ/)	disreputable (/ l̥ ʉ/)	conflagration (/əgr/)	correlation (/e^{y}š/)
flamboyant (/bɔ^{y}/)	locale (/o^{w}ʉ/)	cordwainer (/ɝʉ/)	Eurasian (/e^{y}ž/)
postlude (/ud/)	miscue (/uʉ/)	disreputable (/r ɛp/)	everyday (/e^{y}ʉ/)
loathly (/o^{w}ð/)	rau wolfia (/ʊf/)	everyday (/ɛvr/)	dillydally (/dæ/)
promiscuously (/uʌ/)	skew (/uʉ/)	extraordinarily (/ɝʌ/)	dossier (/i^{y}e^{y}/)
unctuous (/uʌ/)	stalwart (/æw/)	fibroid (/ɑ^{y}br/)	flamboyant (/ɔ^{y}ɛ/)
wallflower (/ɑ^{w}ɝ/)	wallflower (/ɔf/)	horology (/ɝa/)	fraudulent (/ǰɪ/)
	unilateral (/ l̥ ʉ/)	laceration (/ɝe^{y}/)	gladiolus (/i^{y}o^{w}/)
	view (/uʉ/)	luxurious (/ɝɪ/)	Ghanaian (/e^{y}ɛ/)
	walrus (/ɔr/)	periwig (/ɝɪ,')	guarantee (/gæ/)
	whitlow (/o^{w}ʉ/)	plurality (/ɝæ/)	humiliate (/i^{y}e^{y}/)
	withdraw (/ɑ^{w}ʉ/)	urethra (/ʌʉ/)	mansuetude (/tu/)
	wolfram (/ʊf/)	unilateral (/ɝə/)	radiology (/i^{y}ɑ/)
	wolverine (/ʊv/)	wallflower (/ɝʉ/)	radiology (/ǰi^{y}/)
		whippoorwill (/ɪpɝ/)	reconstruct (/kɪ/)
		wolverine (/ɝi^{y}/)	requiem (/i^{y}ɛ/)
			riyal (/i^{y}ɑ/)
			wagonette (/gɪ/)

MAY 22 1993