Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment

by

James Robert Glass
B.Eng., Carleton University
(1982)

Submitted in Partial Fulfillment
of the Requirements for the
Degrees of

Master of Science

and

Electrical Engineer

at the

Massachusetts Institute of Technology

December 1984

Signature of Author ....James R. Glass........
Department of Electrical Engineering and Computer Science
December 21, 1984

Certified by ....V. W. Zue........
Victor W. Zue
Thesis Supervisor

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

Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment

by

James Robert Glass

Submitted to the Department of Electrical Engineering and Computer Science on December 21, 1984 in partial fulfillment of the requirements for the degrees of Master of Science and Electrical Engineer.

Abstract

This thesis is concerned with the acoustic analysis of nasal consonants and nasalized vowels, and the design, implementation, and evaluation of a set of algorithms to detect nasal consonants and nasalized vowels from the speech waveform. The acoustic study uses a database consisting of over 1200 words, excised from continuous speech, and recorded from six speakers, three male and three female. All of the recorded words were digitized, and their phonetic transcriptions were aligned with the speech waveform. Using the Spire and SpireX speech analysis tools, acoustic features common to all nasal consonants and nasalized vowels were determined. For nasal consonants these included the presence of a low frequency resonance in the short-time spectra, centered between 200 and 350 Hz, the global, and local strength of this peak, and a measure of spectral stability in the low frequency regions. For nasalized vowels these included the presence of an extra resonance in the short-time spectra, the relative amplitude of this peak to that of the first formant, and a measure of the broadness of the spectral peak in the first formant region.

The nasal consonant detection algorithms were designed to discriminate between nasal consonants and impostor sounds such as liquids, glides, or voice bars. The nasalized vowel algorithms were designed to discriminate vowels adjacent to a nasal consonant from vowels in other contexts. In each case, a log likelihood decision strategy, using robust measures established in the acoustic analysis, was employed. The detection systems were evaluated on the database by training on the speech of five speakers, and then testing on the tokens of the final speaker. The results indicate that a nasal consonant can be detected 88% of the time, while a vowel adjacent to a nasal consonant can be identified 74% of the time.

Thesis Advisor: Victor W. Zue
Title: Assistant Professor of Electrical Engineering and Computer Science

Acknowledgments

I am extemely grateful to my advisor, Victor Zue, for giving me the opportunity to learn from his expertise in speech analysis and speech recognition. I thank him for his guidance, especially during the difficult times, and for his advice during the preparation of this thesis.

I also thank the members of Zues Crew, and the Speech Communication Group, for many helpful suggestions and discussions, and for creating a stimulating and enjoyable environment in which to conduct research. I thank Mark Randolph in particular, for assisting me with the use of his binary tree classifier, and Scott Cyphers, who spent countless hours maintaining the laboratory Lisp Machines, and improving the Spire facility.

I thank Lori Sorayama, for carefully reviewing drafts of this thesis, and for her support and encouragement for as long as I have known her at MIT.

Finally, I thank my parents, and my brother and sisters, for their continuing love, and encouragement, and for showing me the value of an education.

This research was supported by the Natural Sciences and Engineering Research Council of Canada, and by the Office of Naval Research.

Contents

Chapter 1

Introduction

1.1 Machine Recognition of Speech

The topic of automatic speech recognition has intrigued scientists and engineers for many years. Apart from the brief period of large scale effort in continuous speech recognition witnessed during the ARPA project [28], the majority of research in this field has been directed towards isolated word recognition. This is particularly true of the recent past, where the primary focus of attention has been on the development of small-vocabulary, speaker-dependent, isolated-word recognition systems. These systems tend to be based on general pattern matching techniques and incorporate little speech specific knowledge [25], [34].

Although general pattern matching algorithms excel within their limited problem space, the extension of these techniques to more difficult tasks involving multiple speakers, large vocabularies, or continuous speech have largely been met with limited success. These results have caused many researchers to believe that large recognition systems would be more successful, if they incorporated a better understanding of speech sounds. This belief is reinforced, at least in part, by a series of spectrogram reading experiments by Cole et al, which indicated that the acoustic signal is rich in phonetic information [7]. These experiments revealed that a trained subject, using explicit acoustic-phonetic rules, could phonetically transcribe unknown sentences from speech spectrograms with an accuracy of 85%. This result suggests that automatic phonetic recognition performances have the potential to be substantially better than are presently reported [28].

One of the most important factors leading to this benchmark performance in spectrogram reading was an improved understanding of the acoustic characteristics of fluent speech. Although there has been a significant amount of research over the last forty years on the acoustic properties of speech sounds, little attention has been given to the acoustic characteristics of speech sounds in continuous speech. Over the last decade this has slowly been changing. As Zue has illustrated, we now have a much better understanding of the properties of speech sounds in different phonetic environments [78]. However, there is still a need for basic research directed towards the *quantification* of the acoustic characteristics of speech sounds.

The research in this thesis is motivated with this requirement in mind. The primary objective of this work is to characterize, and quantify, the acoustic properties of nasal consonants and nasalized vowels in American English. Nasal consonants were chosen because they appear to cause difficulty for some speech recognition systems, yet have not been studied as extensively as many other speech sounds. Nasalized vowels were included because of clear indications that they provide important acoustic information about the presence of a nasal consonant.

Once the characteristics of nasal consonants and nasalized vowels are quantified, automatic detection systems, which incorporate robust acoustic measures of nasality, are designed for use in a speaker-independent, continuous-speech environment. Evaluation of these systems provides an indication of their potential for use in speech recognition.

1.2 Acoustic Studies of Speech

1.2.1 The Nature of Speech Sounds

All languages appear to consist of a finite number of distinguishable, mutually exclusive sounds which are concatenated together in time to produce speech. These basic linguistic units are called *phonemes*, and possess unique articulatory and acoustic characteristics [14]. In American English, there are approximately 42 phonemes, which include vowels, semivowels, and consonants [13].

It has long been proposed that there are underlying invariant acoustic properties for all phonemes, which allow an utterance to be decoded from the acoustic signal [26]. However, there are many factors which can influence the observed acoustic pattern of phonemes, and therefore complicate a study of their properties. These factors include:

- *Contextual differences.* When phonemes are connected together to form larger linguistic units, the acoustic characteristics of a given phoneme are modified by the immediate phonetic environment. Occasionally, a speaker can distort the acoustic properties so severely that the phoneme may not be identified, despite a knowledge of the phonetic environment [76]. These distortions are possible because, in addition to acoustic-phonetic knowledge, listeners are able to apply syntactic, semantic, phonetactic, and phonological constraints to help recognize an utterance.

- *Inter-speaker differences.* The acoustic characteristics of speech sounds depend upon the physiological structure of the vocal apparatus which varies from speaker to speaker. In particular, there can be large acoustical differences in the speech of men, women, and children.

- *Intra-speaker differences.* The same speaker can pronounce an utterance differently on separate occasions for many reasons including sickness, mood, audience (e.g. child versus adult), stress patterns on the word or phrase, and transmission environment.

In order to compensate for these factors, many studies, including this one, base their analysis on a carefully designed database. A discussion of the motivation for utilizing databases, and a description of the particular database used in this analysis, are presented later in more detail.

1.2.2 Production of Nasal Sounds

The basic production mechanisms of nasal consonants and nasalized vowels have been studied extensively and are well understood [13], [14]. Nasal consonants are considered to be voiced, since during their production the vocal tract is excited by vocal fold vibration. Nasal consonants are produced by lowering the velum so that air flows through the nasal tract and is radiated at the nostrils (figure 1.1 shows a cross-section of the human vocal apparatus). The closed oral cavity and the sinuses of the nose form shunting cavities to the main path (pharynx and nasal tract) which substantially influences the resulting radiated sound. Figure 1.2 illustrates typical vocal tract configurations for /m/, /n/, and /ŋ/, the three nasal consonants produced in American English. Note that the main difference between the three consonants is the location of the constriction formed with the tongue. Figure 1.3 contains spectrograms of the words *simmer, sinner,* and *singer.*

Nasalized vowels are produced in a similar manner to nasal consonants, with the exception being that the oral cavity is not blocked, thereby allowing air to flow through *both* the nasal and oral cavities.

In many languages, including American English, nasal consonants can have a profound effect on neighboring vowels. Following the release of a nasal consonant, the initial portion of a following vowel will be nasalized during the time interval that the velum is closing. The same holds true for the final portion of a vowel

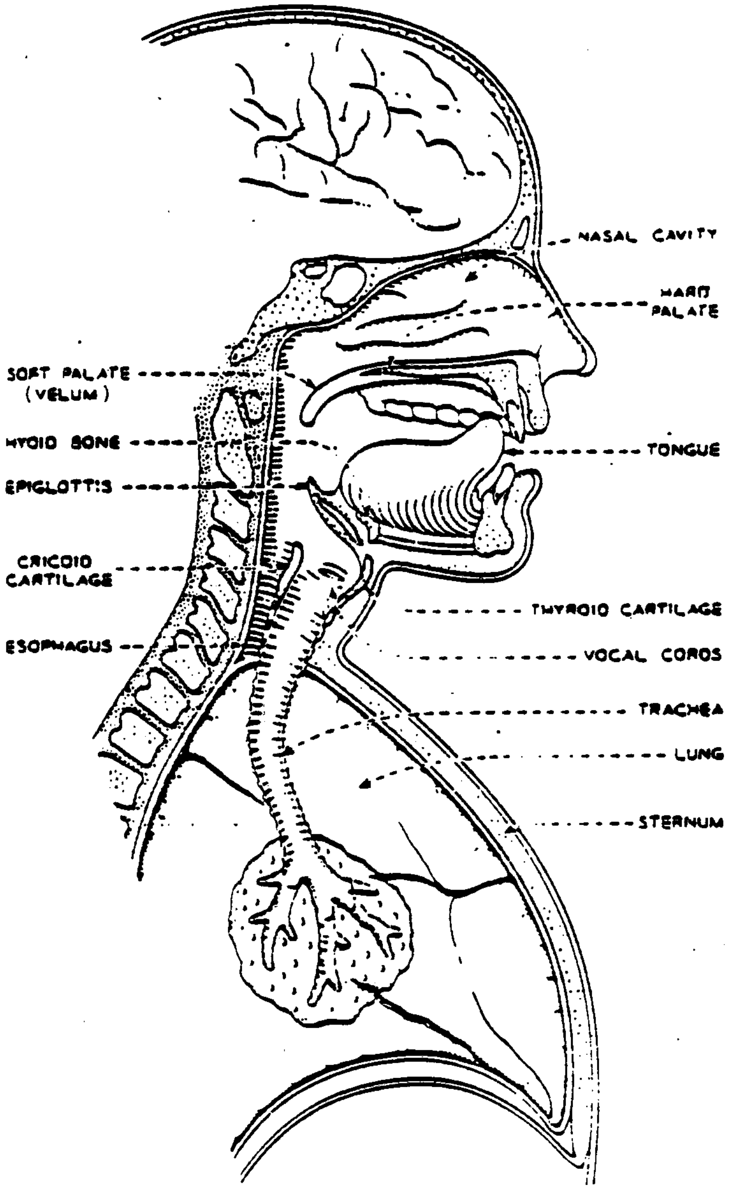


Figure 1.1: The Human Vocal Apparatus (from Flanagan)

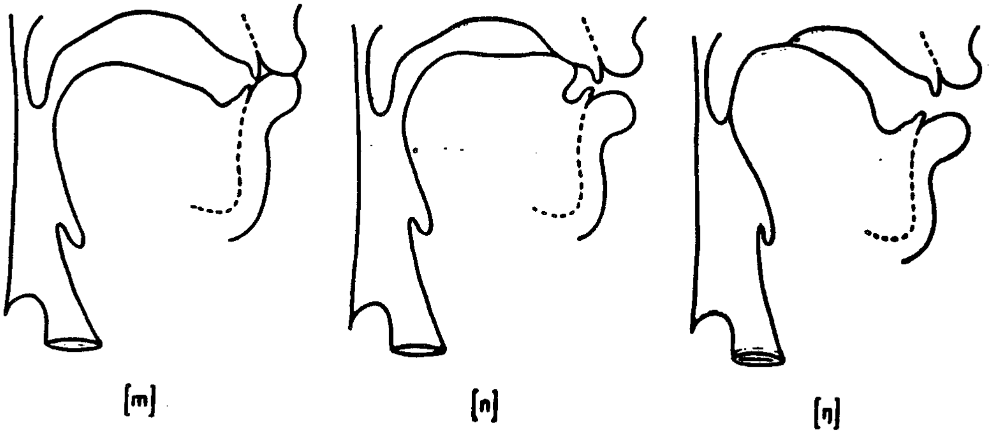


Figure 1.2: The Vocal Tract Configurations of the Nasal Consonants

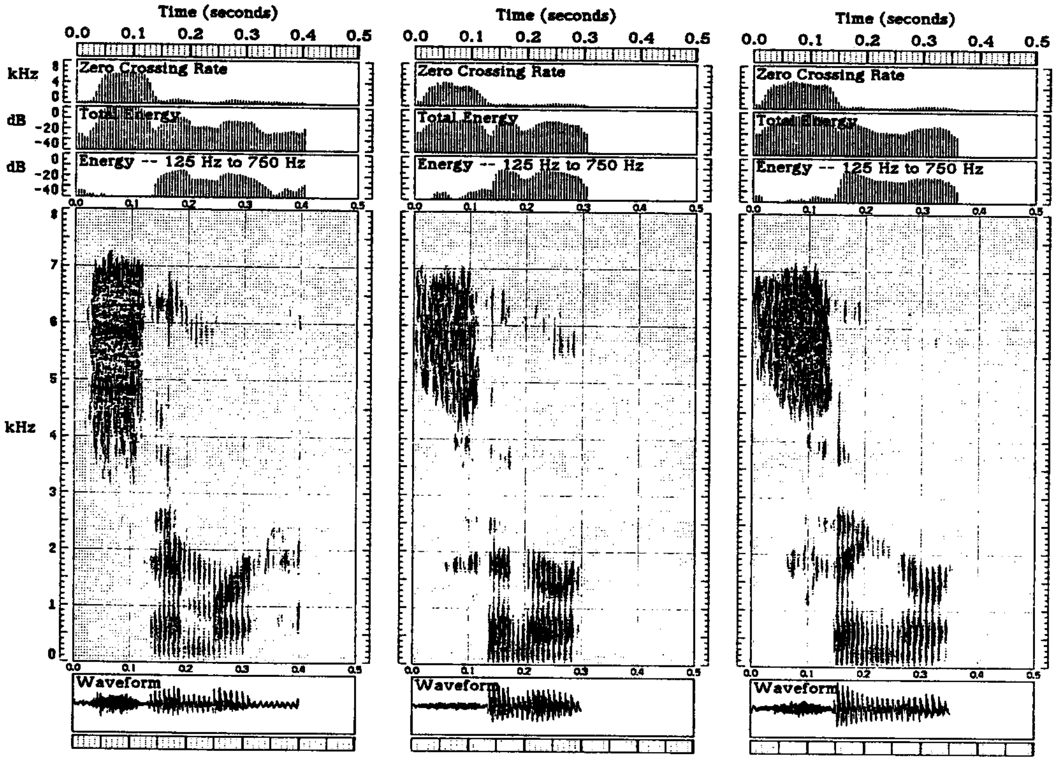


Figure 1.3: Spectrograms of the words *simmer, sinner,* and *singer*

preceding a nasal consonant [21]. The amount of coarticulated nasalization depends upon the particular language and dialect. Since anticipatory nasalization is common in American English [27], a sequence of a vowel plus a nasal consonant (VN) may, in many situations, be pronounced as a simple nasalized vowel, or a nasalized vowel plus a short, residual nasal murmur. This is especially true of vowel-nasal-consonant (VNC) sequences where the consonant is a voiceless stop, as in the words *camp, bent*, or *bunk* [44]. In these cases, nasalization of the preceding vowel may provide the major acoustic difference between these words, and the corresponding pair words *cap, bet*, and *buck*.

In American English, nasalized vowels are not distinguished phonemically from non-nasalized vowels. Thus, speakers have the freedom to nasalize vowels at will, *independent of the presence or absence of a nasal consonant*. For this reason, it is important not to assume that the presence of a nasalized vowel will always indicate the presence of a nasal consonant as well. This research determines if the *relative* degree of vowel nasalization is a more robust indication of the presence, or absence, of a nasal consonant.

In addition to the potential benefit to nasal consonant detection, an understanding of the acoustic characteristics of nasalized vowels would be very useful for many speech analysis tools such as formant trackers, which have traditionally had difficulty with nasalized vowels. Knowledge of the nasalized portions of an utterance would allow formant trackers to employ different, and more successful, strategies in these regions.

1.2.3 Previous Studies of Nasal Sounds

There is a vast amount of literature spanning over twenty-five years which involves the analysis, synthesis, perception, and recognition of nasal consonants and nasalized vowels. The following sections attempt to provide a brief summary of some of this work in order to put the acoustic study of this research into better perspective.

Analysis and Synthesis Studies

There has been a large amount of work which has studied the acoustic characteristics of nasal consonants and nasalized vowels. Much of this research has involved the use of synthetic speech. The following paragraphs summarize some of past work on nasal consonants.

- Using an analog vocal tract model, House found that synthetic nasal consonants were characterized by a predominance of low frequency energy, low overall level compared to a vowel, and a spectral prominence near 1000 Hz [24].

- Fujimura reported several studies of the acoustic characteristics of nasal consonants [15], [16]. He found that the nasal murmur spectra is characterized by the existence of a very low first formant, located at about 300 Hz, which is well separated from the upper formant structure. He also noted that the formants were highly damped, and that there was a high density of formants compared to vowels. Further, he observed the presence of an antiformant, caused by the closure in the oral cavity, which varied in frequency with the place of articulation. In general, the antiformant could be found between 750 and 1250 Hz for /m/, between 1450 and 2200 Hz for /n/, and above 3000 Hz for /ŋ/. He noted however, that although the antiformant varies with the place of articulation, the overall spectral shape of nasal consonants are very similar in appearance.

- From sweep-tone measurements of the vocal tract, Fujimura and Lindqvist reported that the primary characteristics of nasal consonants were a marked, but not necessarily simple, low-frequency boost around 200 to 300 Hz [18].

They also noted a gross deviation from vowel spectral shapes, and a higher total energy compared to stops, both in the low-frequency boost, and in other frequency ranges. They also observed that the transfer function characteristics of the nasal consonants varied greatly from subject to subject.

- Based on evidence from sweep-tone data of the transfer function of the nasal tract, Lindqvist and Sundberg proposed that the complex pole zero patterns observed in nasals and nasalized vowels could be explained by the shunting effect of the sinus cavities [37].

The following paragraphs summarize past work with nasalized vowels.

- House and Stevens studied nasalized vowels with the use of an analog vocal tract synthesizer [23]. They found that the major characteristics of nasalization were a weakened, and broadened first formant, and an overall weaker vowel level than in non-nasalized vowels. They also observed additional weak spectral peaks which tended to fill in the valleys between formants.

- Hattori, Yamamoto, and Fujimura determined that the principal characteristics of nasalization were the presence of a dull resonance around 250 Hz, an antiresonance at about 500 Hz, and additional weak and diffuse components which filled in the valleys between formants [20].

- Fujimura and Lindqvist concluded that nasalization introduces nasal formants into the speech signal [18]. They found that each nasal formant was paired with an antiformant. Depending on the degree of coupling, the antiformant could be either close to the nasal formant, or a nearby oral formant. They also found that as nasalization increases, all formants shifted monotonically upwards.

- Maeda found that by including a model of the sinus cavities, he was able to synthesize a low resonance below the first formant [39]. The addition of this resonance was found to produce natural sounding nasalized vowels of all height.

All of these studies have contributed to the current understanding of the acoustic characteristics of nasal consonants and nasalized vowels. Despite these numerous acoustical studies however, the results are not always directly relevant to speech recognition. Reasons for this include the fact that the data has not been presented in sufficiently quantitative form, or has been presented in relative as opposed to absolute terms. More seriously for automatic speech recognition, some of the data has been obtained from displays where a human must make an interpretation to make a measurement. Finally, in many cases, the data has been obtained from restricted environments such as stressed, consonant-vowel (CV) syllables.

Further insights could be obtained by performing an analysis on a body of naturally spoken data. By providing a better understanding of the variability of the acoustic characteristics of nasal consonants and nasalized vowels in a natural speech environment, such a study would be valuable to scientists concerned with the automatic detection of these sounds.

Perceptual Studies

Much perceptual research has been devoted to studying the role of the nasal murmur and an adjacent vowel, in determining both the manner and place of articulation of the nasal consonant. This section summarizes the results of several of these studies.

- The work of Malécot, Nakata, Nord, Recasens, Kurowski and Blumstein, and Repp, has been concerned with the relative importance of the nasal murmur, and the formant transitions in adjacent vowels, to the identification of the place of the nasal consonant [43], [50], [53], [63], [31], [64]. The common conclusion was that formant transitions were the major cue to place for prevocalic nasals, while for post-vocalic nasals, the murmur was taken into account as well. The work of Kurowski and Blumstein, and Repp, found that the nasal murmur was more informative in utterance-initial position than did previous studies however.

- Malécot has also done perceptual work with homorganic nasal stop consonant clusters [44]. He found that the short nasal murmur played a very minor role in conveying the impression that a nasal was present. The nasalized vowel appeared to be the major cue to the presence of the nasal consonant.

- Mártony reported studies on synthetic nasal production which indicated that damping in the second formant region was very important for natural sounding NV tokens [46]. He also found that the bandwidth values for the nasal murmur of /m/ were much more vowel dependent than /n/ since in /m/ tends to be coarticulated with vowels more than /n/.[1]

- Ali et al reported an experiment indicating that subjects are able to predict the presence of a nasal consonant from the preceding vowel [1]. They hypothesized that listeners use the anticipatory nasalization feature, common for nasal production in English, to help lighten the phoneme processing load.

- Lintz and Sherman investigated the effect of different consonants on the perceived nasality of vowels in CVC tokens [38]. They found that low vowels were judged more nasal than high vowels, front vowels more nasal than back vowels. They also found that nasality is least severe for voiceless plosive environments, more severe for voiceless fricative and voiced plosive environments, and most severe for voiced fricative environments.

- Kawasaki found that vowels in NVN tokens were considered more nasalized when nasal murmur amplitudes were decreased relative to the vowel

[1] In American English the tongue has no distinctive function for /m/, unlike for /n/ or /ŋ/. Therefore, /m/ tends to be coarticulated with adjacent vowels much more than other nasal consonants [72].

amplitude [27]. She also noted that playing the speech backwards made vowel nasalization much more apparent, and attributed this to the fact that listeners do not expect significant perservatory nasalization in English.

- Hawkins and Stevens have reported a perceptual study which indicates that the basic acoustic property of nasalization is a reduction in the degree of prominence of the first formant peak [21]. This reduction is realized by splitting or broadening the first formant spectral peak by creating an additional spectral peak nearby.

Perceptual studies have provided information about the role of different acoustic characteristics in establishing the property of nasality. From a speech recognition perspective, it would be useful to determine, based on acoustic information alone, the inherent recognizability of nasal consonants and nasalized vowels in American English. In other words, listeners would be allowed to use only their acoustic knowledge to decide on the feature nasal; syntactic, semantic and phonetactic information would, as much as possible, be eliminated. This test would provide two benefits. First, it provides an upper bound on automatic recognition performance in the same circumstances. Second, it provides a means of evaluating the perceptual relevance of a set of acoustic characteristics; a high correlation between acoustic measurements and perceptual score being the evaluation measure.

Recognition of Nasal Consonants

There have been several attempts at automatic recognition of nasal consonants:

- Gillmann reported his attempts at nasal identification for post-vocalic nasal consonants considering only the formants in the nasal murmur (by picking peaks of LPC spectra) [19]. He found that the formants did not change appreciably during the murmur, and that formant frequencies were fairly stable for one speaker, although they varied from speaker to speaker. There were enough differences between nasal formant values for any one speaker that he was able to achieve 70% correct nasal identification using a simple least squares clustering procedure.

- Formant frequencies were also used to detect nasal consonants in the sonorant regions of the acoustic-phonetic analysis system developed by Weinstein et al. [73]. Nasal consonants were required to pass a duration constraint, as well as speaker dependent constraints on the formant values (such as low value of the first formant frequency, low ratio of second formant amplitude to first formant amplitude, and a higher ratio of third formant amplitude to first formant amplitude) to be accepted. They found that nasals were detected correctly about 80% of the time, with intervocalic nasals being detected much more reliably than non-intervocalic nasals. Prevocalic nasals were detected 80% of the time, while post-vocalic nasals were detected only 60% of the time due, in their opinion, to the reduction of the nasal murmur lengths in some environments. They noted that in these situations the adjacent vowel was quite often nasalized. About 15% of the detected pre- or post-vocalic nasals were the phonemes /l/, /w/, or /r/, and another 20% were false alarms (no segment present), caused by vowels with low first formant frequencies, such as /i/, /e/, and /u/.

- Using the hypothesis that nasal boundaries can be found at points of maximal spectral change, Mermelstein attempted a very ambitious project to detect nasal consonants in continuous speech [48]. He used four simple spectral measurements to classify the region adjacent to these transitions as either nasal or non-nasal. Using a multivariate statistical training procedure, he was able to obtain a 91% correct nasal/non-nasal decision rate on paragraphs spoken by two male speakers. Mermelstein also found that speaker dependent training was superior to speaker independent. He pointed out that the majority of errors confused nasals with weak fricatives and /l/

and /r/ before high vowels. He also pointed out that nasal segments were missed when they were shortened.

- Hess reported a 90% recognition rate for German nasals in continuous speech for a single speaker [22]. Dixson and Silverman reported a 94% recognition rate for nasals in continuous speech for one speaker [11].

- De Mori has reported work on discriminating intervocalic /n/ and /m/ in continuous speech [9]. Decision making was based on the value of the second formant at the beginning and end of the nasal consonant and the amplitude differences between the formants at the point during the nasal murmur where the second formant amplitude is minimal. Tested on four male speakers, the average error rate was 6% with the majority of error occurring in a front vowel environment.

There are two points which can be made about these studies. First, none of these efforts has reported testing their systems on a large number of speakers. The system was either designed to be speaker dependent, or was tested on very few speakers (all male). Clearly, the strong speaker dependent characteristics of nasal consonants present a challenge to any recognition system. In order to claim that a system is speaker independent, it is necessary to test it on a much larger number of speakers. The second point of note is that there have not been many, if any, attempts to automatically detect nasalized vowels, even though researchers have noted that this capability would be very beneficial to help verify the presence of a nasal consonant.

1.3 Summary and Outline of Research

There is clear evidence that the acoustic signal of speech is rich in acoustic information. This implies, that by incorporating more knowledge about the acoustic characteristics of speech sounds, automatic phonetic recognition performances have the potential to be substantially better than are presently obtained in practice.

A survey of previous acoustic studies reported in the literature indicates however, that while the results clearly establish relevant acoustic properties of the speech sound, they are not always directly applicable to speech recognition systems, due to the manner in which measurements were calculated, or due to the nature of the analysis database itself.

The primary objective of this thesis research is the characterization and quantification of nasal consonants and nasalized vowels in American English. The secondary objective is to design automatic nasal consonant and nasalized vowel detection systems which incorporate robust acoustic measures of nasality, and operate in a speaker-independent, continuous-speech environment.

The research in this thesis is organized into two stages. First, an acoustic study of nasal consonants and nasalized vowels is conducted. The main goal of this study is to observe, and quantify the observations made by previous studies, using a large database of natural utterances. Chapter two describes the methodology used for the acoustic study, and chapter three presents the results of the data analysis.

The second stage of this research is concerned with the automatic detection of nasal consonants and nasalized vowels in continuous speech. Chapter four describes and evaluates the detection systems, and reports on a set of experiments designed to determine the perceptual merit of the system decisions.

Chapter five presents a summary of the thesis. In addition, suggestions for further research are discussed.

Chapter 2

Data Analysis Methodology

The acoustic analysis of nasal consonants and nasalized vowels, is performed through a series of experiments, and is conducted on a database of utterances. The design of the database requires that several important issues be considered. These issues are discussed in the next section along with a description of the database construction. The following section describes the data analysis procedures used in the acoustic study, and the final section briefly describes the data analysis facility used for all of the acoustic experiments.

2.1 Database Description

Due to the variability of the speech waveform, any attempt to quantify acoustic characteristics of speech sounds requires a carefully designed database. In the past, the majority of researchers have opted to study speech sounds in restricted environments, such as stressed consonant-vowel sequences embedded in nonsense syllables. The theory behind this methodology is that stressed syllables are probably articulated with greater care and effort, and thereby produce a robust acoustic signal whose features may be extracted more reliably [70], [71].

A study using naturally spoken words however, provides greater insight into the acoustic characteristics of sounds in fluent speech. Also, any quantified observations are more useful for automatic continuous speech recognition, since they give a better indication of the variability of these acoustic characteristics. For these reasons, the database was constructed from real words spliced out of continuous speech.

Once the decision was made to construct the database from naturally spoken words, it became necessary to decide which words to include in the corpus. Since the size of the corpus should be as compact as possible, it was important to create one that was well balanced. Thus, the corpus was created using the following criteria:

- The corpus should contain a diverse sampling of the many possible syllabic and phonetic contexts of nasal consonants in American English. For example, the corpus should contain nasal consonants in intervocalic, post-fricative, and homorganic nasal stop consonant environments as found in the words *conic, smack*, and *pink* respectively.

- The corpus should contain minimal pairs (tokens which have only one phonetic difference), in order to distinguish which acoustic characteristics belong to the nasal consonant class, and which ones do not. For example, the corpus should contain minimal pairs which differ only by the absence of a nasal consonant, as is found in the words *bent* and *bet.* The corpus should also contain minimal pairs which differ only by the substitution of a similar speech sound for the nasal consonant (such as a glide or a voice bar), as is found in the words *made*, and *bade* or *wade.*

- The corpus should contain minimal pairs which can be used to detect acoustic differences within the nasal consonant class itself, such as in the words *simmer, sinner* and *singer.*

- The corpus should contain minimal pairs which can be used to establish acoustic differences between nasal consonants in a poly-syllabic versus mono-syllabic environment. Consider for example, the words *meat* and *voltmeter*, where the syllable-initial nasal consonant has gone from a primary to a secondary stress position.

The contents of the over 200 word corpus may be found in Appendix A.

To produce a database containing utterances which are truly "naturally spoken", the corpus words should be embedded in sentences with acceptable semantic and syntactic structures. However, this type of recording procedure, besides creating a requirement for a large number of carrier sentences, raises the issue of the effect of local syntax and semantics on the individual words. Accounting for this variability would be difficult with different carrier sentences. For this reason, a common carrier sentence was used for all words, so that the corpus words would always be in the same context in the sentence. In this research, the carrier phrase "*She said* ___ *happily*" was used, since it minimized the amount of coarticulation with any word-initial, or word-final nasal consonant, since the phoneme /h/ is neutral, and the phoneme /d/ cannot form a consonant cluster with word-initial nasal consonants in English.

Recordings were made in a sound-isolated room using a Sony omni-directional, electret microphone (model ECM-50PS), a Shure microphone mixer (model M68FC), and a Nakamichi LX-5 tape recorder. The overall signal-to-noise ratio was approximately 30 dB. Original utterances were stored on cassette tape (TDK SA-C60). For recording purposes, the corpus words were randomized into groups of ten. During the recording sessions, speakers were instructed to read naturally and to take a breath at the ends (as opposed to the middle) of phrases. Speakers were allowed to pause for as long as they wanted between each group of ten, but were asked to read each group continuously. Any mispronounced words were repeated immediately following a group of ten. After the recording session, the first and last utterances from each group of ten were deleted in an attempt to minimize artifacts which can occur at the beginning and end of paragraphs.

The analysis database was made from six native speakers of American English (three male and three female) between the ages of twenty and forty. All 1200 utterances were digitized at 16 kHz/s (16 bit words), and their phonetic transcriptions were manually time aligned with the waveform. The time alignment procedures used for the transcription process are described in detail in Appendix B.

2.2 Data Analysis Procedures

The data analysis of nasal consonants and nasalized vowels are divided into three separate studies of duration, energy, and spectral properties. The following sections elaborate on the types of measurements made in each area, and explain how the calculations are computed.

2.2.1 Analysis of Duration

The primary focus of the durational study is to quantify the effect of phonetic context on the duration of the nasal consonant. Nasal consonant durations have been studied more than any other acoustic characteristic of the nasal consonant. Thus, it is easy to compare the results of previous work to those found in this study. Many studies in the past have restricted themselves to one particular phonetic context, such as homorganic nasal stop consonant clusters. An important contribution made by this work therefore, is to allow a comparison of nasal consonant durations in many different environments.

The duration of nasalized vowels are quantified in order to observe their durations relative to oral vowels. Once again, a comparison will be made with previously reported results.

Calculation of Duration

Duration is relatively simple to compute, since it is defined by the time alignment of the phonetic transcription. Although in the past, it has not always been an easy matter to find the exact boundaries of any given phoneme [44], the use of spectrograms simplifies this task. For instance, the temporal boundaries of the nasal consonant are relatively easy to establish, since they are usually denoted by sharp spectral changes which occur at the beginning and end of the period of oral closure. In general, boundaries produced by different transcription experts are within 10 msec of each other [35].

2.2.2 Analysis of Energy

There are two procedures used to analyze the energy characteristics of nasal consonants. Since nasal consonants occur next to a vowel in English, the first procedure measures the relative difference in average energy between the nasal consonant and an adjacent vowel. When the nasal consonant occurs in a medial context, the largest energy difference is computed.

From a speech recognition perspective, it would be valuable to know how the distribution of this energy difference of nasal consonants compares with other sounds. This would establish if the energy difference measure has any potential for use in a discrimination task. Thus, comparisons are made to sounds with similar acoustic characteristics to nasal consonants, such as semivowels and voice bars. Figure 2.1 contains spectrograms of the words *hammock, cab,* and *lip.* Note that the semivowel /l/, and the voice bar in /b/, have similar acoustic characteristics with the /m/.[1]

[1] Since voice bars are not always immediately adjacent to a vowel, a slightly different procedure was also used to quantify energy. In this case, the energy value in the token is relative to the largest energy in the utterance instead of an adjacent vowel. Both of these procedures were found to produce similar results.

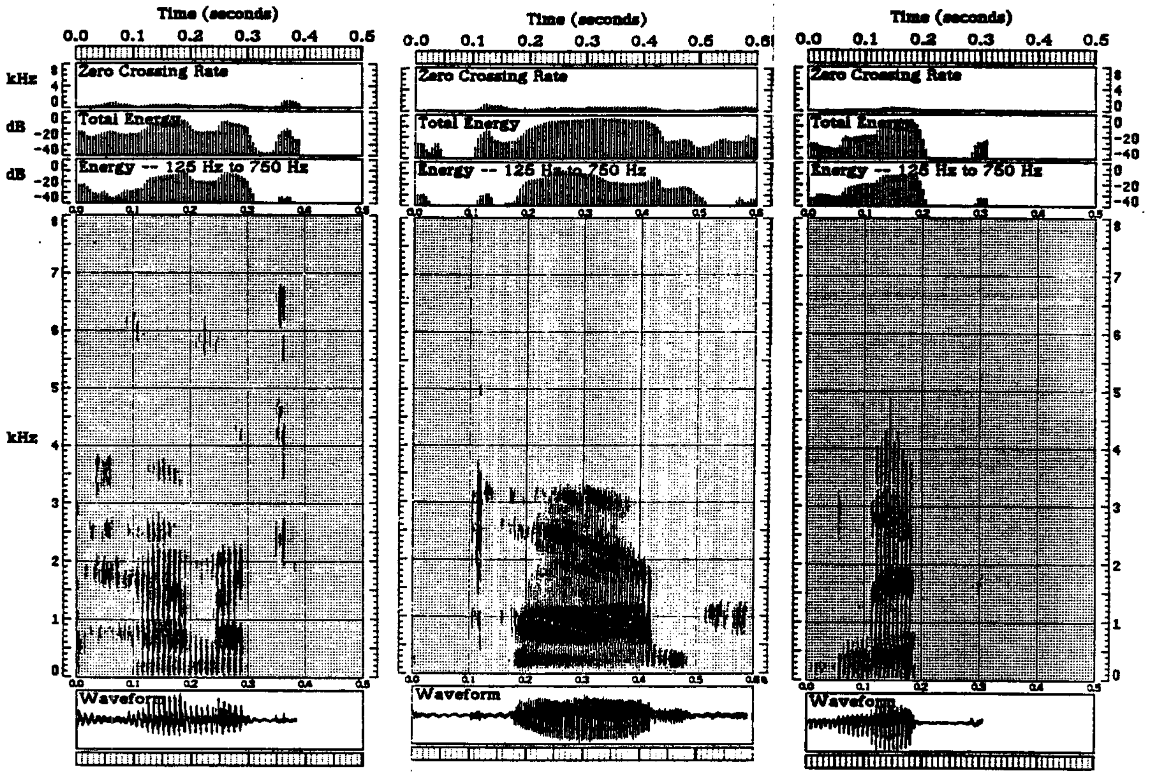


Figure 2.1: Spectrograms of the words *hammock, cab,* and *lip*

As nasal consonants are commonly believed to be stable, since their vocal apparatus is held fixed during production, it is of interest to measure how much the energy parameters actually change during the nasal murmur. The stability is measured by computing parameters such as standard deviations of energy values, and average values of first differences of energies in the nasal murmur. Once again, comparisons are made between nasal consonants and similar speech sounds.

Calculation of Energy

Energies are calculated using short-time processing techniques commonly used in digital speech processing. The underlying assumption for the use of these procedures, is that the vocal mechanism is quasi-stationary, in that its acoustic characteristics change slowly with time. Thus, short segments of the speech signal may be isolated and processed as if they were short segments from a sustained sound. In general, the short-time energy is defined as

$$E_n = \sum_{m=-\infty}^{\infty} (x[m]w[n-m])^2 \tag{2.1}$$

where $x[n]$ is the speech waveform, and $w[n]$ is a windowing filter, the shape of which can drastically affect the short-time energy function E_n. In general, it is desirable to have a window with an impulse response short enough so that the energy function is responsive to rapid changes in the speech signal. However, the impulse response should also be long enough to provide sufficient averaging of the speech waveform to produce a smooth energy function. Further discussions on windowing may be found in speech processing textbooks [56]. For many digital speech processing applications, a hamming window is used [61]. In this research, a hamming window of 25 msec duration was used in all of the energy calculations.

Energy in a particular frequency band is computed by taking the dot product of the short-time spectra, $X(e^{j\omega})$, with a frequency window, $Z(e^{j\omega})$, typically of trapezoidal shape (Appendix C contains a discussion of short-time Fourier spectra). Using Parsevals relation for conservation of energy, it can be shown that this procedure is equivalent to producing the short-time energy via equation 2.1 when $x[n]$ is first filtered by a function with frequency response $Z(e^{j\omega})$.

During data analysis, all energies were converted to dB to reduce the sensitivity of the energy function to small changes when the energy signal is large.

For the statistical analysis of energy stability, the average energy of a token, $\bar{E}$, and its standard deviation, σ, computed between two time points n_1 and n_2, are defined as

$$\bar{E} = \frac{1}{n_2 - n_1} \sum_{n=n_1}^{n_2} E_n \tag{2.2}$$

$$\sigma = \sqrt{\frac{1}{n_2 - n_1} \sum_{n=n_1}^{n_2} (E_n - \bar{E})^2} \tag{2.3}$$

2.2.3 Analysis of Spectra

Perhaps the most interesting aspect of the acoustic study is the study of the spectral characteristics of the nasal consonants and the nasalized vowels. The spectral analysis performed in this research is carried out in two steps. In the first stage of analysis, the goal is to establish prototypical spectral shapes. From these spectral shapes, it is possible to hypothesize general spectral characteristics of the nasal consonant or nasalized vowel.

The next step in the analysis is to develop algorithms which are able to automatically extract the properties observed in the prototypical spectral shapes. Due to the variability of the speech signal across speaker and context, the emphasis at this stage is on creating measurements which extract information about *robust* characteristics of the nasal consonant or nasalized vowel. Algorithms which try to measure subtle properties of nasality are often fragile, and sensitive to speaker variability, and hence are avoided wherever possible.[2]

Once measurement algorithms are created, the characteristics of utterances in the database are quantified. As was the case for duration and energy, comparisons will be made between the distributions of nasal consonants and those of similar sounds, and between nasalized and non-nasalized vowels.

Finally, part of the analysis is concerned with measuring the spectral stability of the nasal consonant. While there is clearly a significant spectral change at the transition between a nasal consonant and an adjacent vowel, it is worthwhile to quantify the spectral stability of the nasal murmur itself, and to compare this stability to that of similar speech sounds.

Calculation of Spectra

All spectral analysis is based on smoothed spectra computed with the discrete

[2] The actual algorithms used in the data analysis are described in detail in the following chapter.

Fourier transform (DFT). The spectra were computed every 5 msec, and were smoothed by windowing the cepstra with a low-pass window that is constant for the first 1.5 msec, and cosine tapered for the next 1.5 msec. Figure 2.2 illustrates an unsmoothed and a smoothed version of a DFT, taken from a nasalized /i/ in the word *technique.* A discussion on the issues involved in spectral analysis may be found in Appendix C.

Wherever it is desired to compute parameters based on the smoothed spectra itself, the spectra are sectioned into peaks, valleys, and transition regions through the use of the second derivative of the smoothed spectra. Boundaries are located at zero crossings of the second derivative spectral slice. Figure 2.3 illustrates an example of a spectral slice which has been schematized in this manner. From this point it is easy to establish spectral peaks and valleys.

Although there are no formal procedures, there are several methods which can be used to measure spectral change in the speech signal. Ultimately, each technique attempts to measure some difference in consecutive short-time spectra. One simple method consists of observing changes in the first few cepstral coefficients, since by their definition,

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{j\omega})|e^{j\omega n}d\omega \tag{2.4}$$

they just weight the log spectrum by different shaped cosine windows.

Naturally, there is no reason why the windows cannot be an arbitrarily shaped function. In fact, it is quite often advantageous to shape a window function, $W(e^{j\omega})$, so that it is sensitive to spectral changes in a particular frequency region. One way of computing the spectral change parameter, S, is by taking the normalized dot product of the spectral slice, $X(e^{j\omega})$, with the weighting window, $W(e^{j\omega})$,

$$S = \frac{\vec{X} \bullet \vec{W}}{|\vec{X}||\vec{W}|} \tag{2.5}$$

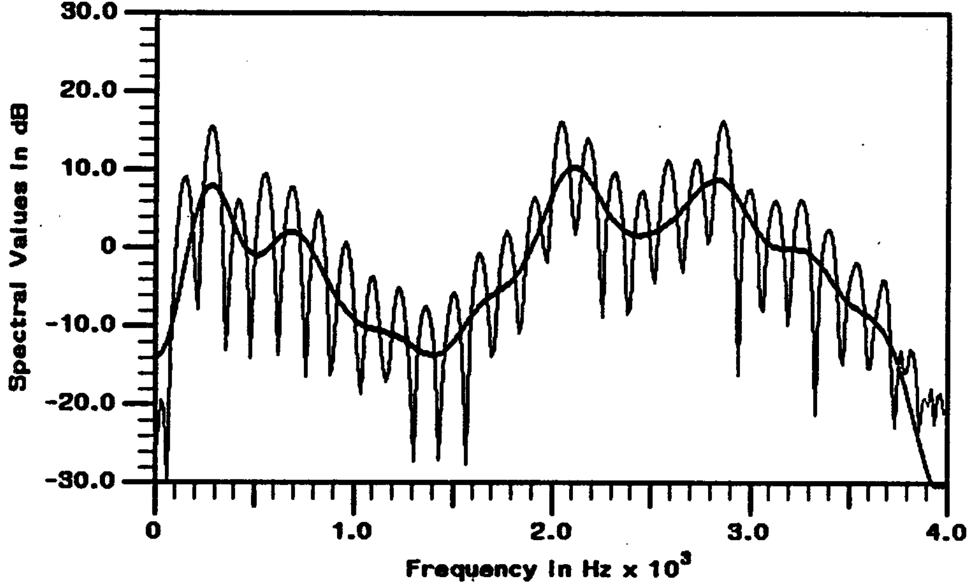


Figure 2.2: An Unsmoothed and Smoothed DFT Spectral Slice

The smoothed DFT spectral slice is computed by windowing the cepstrum with a window that is flat for the first 1.5 msec, and cosine tapered for the next 1.5 msec. This particular example was taken from the /i/ in the word *technique*

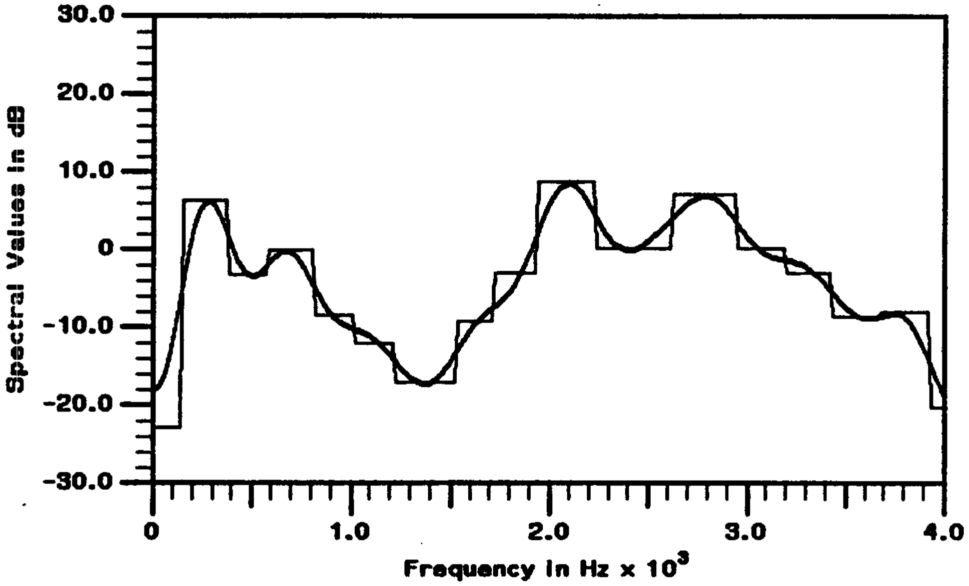


Figure 2.3: A Schematized Spectral Slice

The smoothed DFT spectral slice is schematized into peaks, valleys, and transition regions. Boundaries are located at zero crossings of the second derivative of the smoothed spectral slice.

Note that both of the parameters are treated as vectors in equation 2.5. The magnitude of the spectral change may be computed by taking a first difference of this function.

Spectral Averaging

As previously mentioned, the first stage of the spectral analysis of nasal consonant and nasalized vowel spectra establishes prototypical spectral shapes. Since nasal consonants have little flexibility in the manner in which they are produced, one might expect that for a given speaker, and a given place of articulation, the murmur spectra could be averaged together without a significant loss of information. This argument can be extended to steady state vowels as well. The validity of this procedure is indicated by the size of the variance in the spectral average. There are other, more sophisticated forms of clustering or data reduction such as k-nearest-neighbor, or principal component analysis, which have been used successfully in the past for speech sounds [30], [60], [72]. However, since the objective of this first stage is mainly for qualitative observation, and not for quantification, a more sophisticated analysis procedure is not pursued.

For analysis, spectra are pre-emphasized, and computed from a windowed cepstrum. As well, the spectra are all normalized with respect to total energy so that individual energy offsets are eliminated. Analysis is restricted to one speaker at a time, in order to eliminate speaker variability.

For nasal consonant analysis, statistics are gathered by collecting multiple spectra from all of the nasal murmurs. Figure 2.4 shows multiple spectra for /m/ for a female speaker. Note that there appears to be common characteristics among the many spectra, suggesting that averaging is reasonable in these circumstances. In pilot studies, there were actually two different averaging procedures which were evaluated. Figure 2.5 shows the average spectra obtained from collecting multiple spectra from each nasal murmur, while figure 2.5 shows the average spectra obtained by collecting a single average spectra from each nasal murmur. In the figures, the thick line is the mean spectral shape, and the outer two lines are one standard deviation away. As can be seen, the average spectral shapes are very similar. The standard deviation of the multiple spectra averaging technique is slightly larger. This is to be expected however, since there are a larger number of spectra included in the averaging. The fact that the two averaging techniques yield similar results illustrates that the spectral characteristics of the nasal murmur are quite stable against time, especially at low frequencies. Since both spectral averaging techniques yielded similar results, the multiple spectral averaging procedure was used for all data analysis since it gave a better indication of the variance of the spectral shapes.

The same multiple spectra averaging technique is used for analysis of nasalized vowels. Even though the averaging procedure is quite informative, care must be taken in interpreting the average spectra, since nasalization is not a static spectral characteristic, but often changes the duration of a vowel. Figure 2.6 illustrates the case for the word *mitt*, where the spectral characteristics of the low resonance region on the left side of the vowel, are clearly different from those on the right.

2.3 Data Measurement Facility

Data analysis is performed with the Spire and SpireX facilities available on MIT Lisp machine workstations [69]. SpireX is a statistical analysis package which allows the user to perform acoustic-phonetic experiments on a large body of utterances. Using SpireX, a typical experiment proceeds in five steps, each of which is described in the following paragraphs.

Catalogs

A user first specifies a *catalog* of utterances to be used for the experiment. A

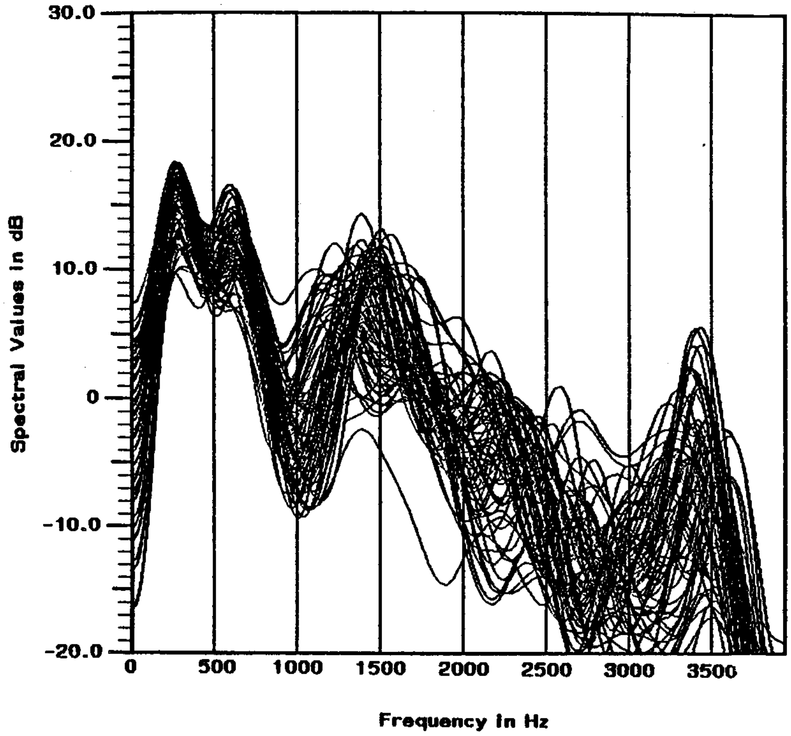


Figure 2.4: Multiple Spectra of an Intervocalic /m/

This display presents an overlay of the normalized smoothed spectra occurring during the nasal murmur of an intervocalic /m/ for a female speaker.

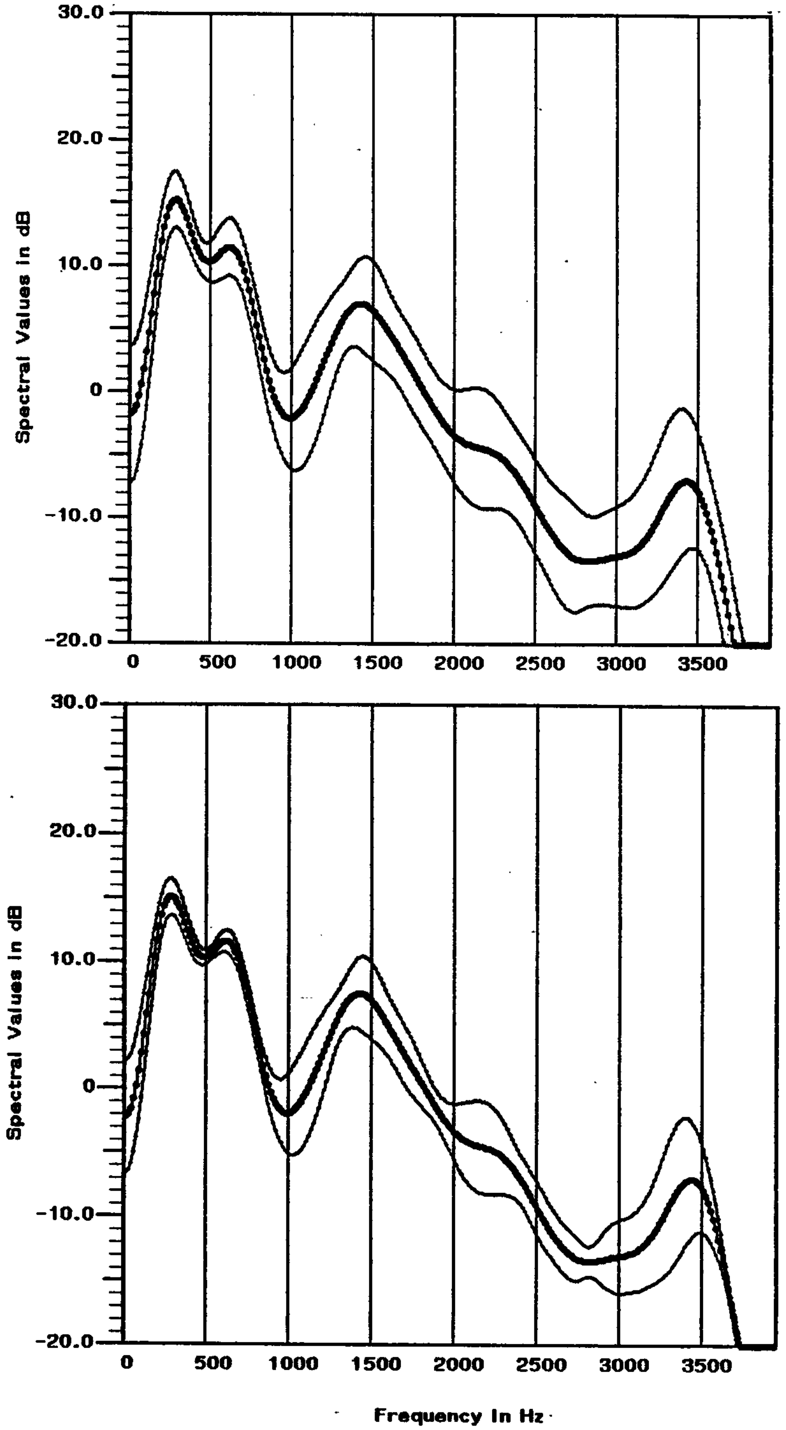


Figure 2.5: Spectra Averaging Techniques

This figure illustrates two techniques which produce a statistical summary of the normalized smoothed spectra of figure 2.4. In the top display, multiple spectra were collected from each murmur. In the bottom display, an average spectra was collected from each murmur. The average spectral shape, shown by the dark line, is surrounded by lines which represent one standard deviation from the mean.

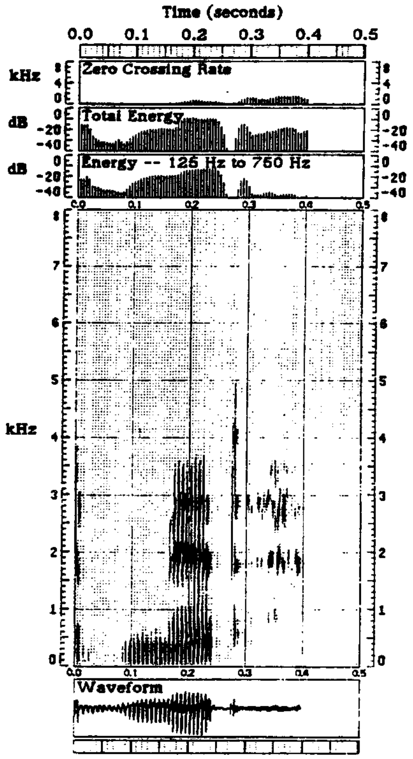


Figure 2.6: A Spectrogram of the word *mitt*

catalog contains information about a set of utterances. In particular, it contains information about utterance filenames, orthographic, and phonetic transcriptions. This information allows SpireX to determine the utterances used in a given experiment without the necessity of loading in each utterance, thus saving both time and memory.

In addition to the transcription information, SpireX catalogs are also able to store attribute values as well. Precomputing attribute values saves on time, since they do not have to be computed during the experiment. Also, memory space is conserved since utterance waveforms, which require a significant amount of space in memory, will not need to be loaded.

Searches

After the catalog has been loaded, and the user has specified a particular phonetic context of interest, SpireX searches through the catalog for instances of the desired phonetic context. Each such instance is known as a *sample.* The phonetic context is specified as a sequence of named regions, each of which consists of a given phonetic pattern. For example, a region could specify a class of phonemes, a specific phoneme, or a more complicated pattern. Thus, to collect a sample set of all nasal stop consonant clusters in a catalog, where the nasal and stop are homorganic and dental, the search specification could consist of a sequence of the three regions *vowel, nasal, stop*, where *vowel* is any vowel, *nasal* is an /n/, and *stop* is a /d/, or a /t/.

Once the search is completed, each region is associated with a time-interval for each sample in the sample set. The region names are used as arguments in later steps of the experiment to reference these time-intervals.

Computations

After the search is complete, the user then specifies a set of computations to be performed on each sample. Computations are usually supplied with search regions and Spire attribute names as arguments. A computation then performs statistical measurements of the attributes in the time-intervals specified by these regions. Typical computations include averages, maximums, and durations. Although computations are usually specified in terms a menu driven interface, users are also allowed to define their own computations, although this requires some knowledge of SpireX. In the nasal stop consonant cluster example, typical computations might include the durations of the *nasal*, and *vowel* regions, and a binary computation which indicates if the *stop* is voiced, or voiceless.

Filters

Filters are logical computations which are used to separate the sample set into groups. Only samples which match a filtering specification are included in the statistical analysis. Thus, for the nasal stop example, the voicing computation could be used to filter the sample set. This would allow the user to separate the statistics of the *nasal*, and *vowel* duration computations of voiced stops from those of voiceless stops.

Display

Once the sample set has been filtered, it is possible to perform a statistical analysis on specified computations and view or tabulate the results. The display capabilities include histograms, scatter plots, and statistical summaries.

2.4 Chapter Summary

This chapter discussed the methodology used for the acoustic analysis of nasal consonants and nasalized vowels. The major points of the chapter were,

1. Data analysis is accomplished by performing a series of experiments on a database of utterances.

2. The database consists of over 1200 words, excised from continuous speech, and recorded from six speakers, three male, and three female. All of the recorded words were digitized, and their phonetic transcriptions were aligned with the speech waveform.

3. Data analysis is divided into a study of nasal consonants and nasalized vowels. In each study, measurements are made of the duration, energy, and spectral characteristics.

4. Data analysis is performed using the SpireX statistical analysis facility.

Chapter 3

Data Analysis

This chapter presents the results of the data analysis experiments carried out on the utterances in the database. The analysis was performed separately on the nasal consonants, and nasalized vowels. The results of each are presented in the following sections.

3.1 Analysis of Nasal Consonants

3.1.1 A Study of Nasal Consonant Duration

Minimal Pair Experiments

As a first step at analyzing the effects of phonetic context on the duration of the nasal consonant, the differences of minimal word pairs, such as *bend/bent*, or *mack/smack*, were observed. The minimal pairs were restricted to monosyllabic words in order to eliminate possible secondary effects introduced in multi-syllable environments (the nasal murmur in the word *picnic* is much shorter than in the word *nick* for instance). The results of these minimal pair experiments, which included all of the speakers in the database, are presented in figure 3.1, and are summarized below:

1. The durations of word final nasals are lengthened when clustered with a voiced stop consonant (VS), such as for the minimal pair *ben/bend.* For the utterances in the database, the average duration increase was 10 msec, or 20% of the duration of the singleton nasal. Also evident from the figure is that word final nasals are shortened when clustered with an unvoiced stop consonant (US), such as for the minimal pair *ben/bent.* The average duration decrease was found to be 20 msec, or 40% of the singleton nasal duration.

2. The same trends are observed when word final nasals are clustered with a fricative. When the fricative is voiced (VF), as in the minimal pair *one/ones*, the average duration increase is 28 msec, or 35% of the singleton nasal duration. When the fricative is unvoiced (UF), as in the minimal pair *one/once*, the average duration decrease is 18 msec, or 30% of the singleton nasal duration.

3. The duration of word initial nasals are shortened when clustered with an unvoiced fricative consonant (F), such as for the minimal pair *nack/snack.* The average duration decrease was observed to be 40 msec, or 50% of the singleton nasal duration. Since there are no word initial voiced fricative nasal clusters in American English, the opposite trend could not be observed.

4. As implied by the previous experiments, the duration of a nasal in a word final consonant cluster is longer when the clustering consonant is voiced, than when it is voiceless. When the difference for stop consonants (VUS), such as for the minimal pair words *canned/can't*, was observed, the average difference in duration of the nasal consonant was 25 msec. Note that only the phonemes /t/ and /d/ were relevant here, since there are no word final nasal stop consonant clusters with the phonemes /b/, and /g/.

5. The same trends were observed in word final nasal fricative clusters (VUF), such as for the minimal pair words *ones/once.* The average difference was measured to be 40 msec.

Two interesting observations were made from these experiments. Using the knowledge that voiced stop consonants have shorter stop gaps than voiceless stop consonants [76], a simple guideline was established for distinguishing voicing in nasal stop consonant clusters, as shown in figure 3.2. It was found that when the nasal murmur occupied over 80% of the duration of the nasal murmur and stop gap, the stop consonant was voiced 90% of the time. If the fraction was less than 0.7 however, the stop consonant was unvoiced 87% of the time. This observation included poly-syllabic words as well.

As shown in figure 3.3, this same observation was found to hold true for stop nasal consonant sequences, such as /pm/, in the word *chipmunk*. When the fraction was less than 0.4, the stop consonant was unvoiced 83% of the time. When the fraction was greater than 0.5 the stop consonant is voiced 88% of the time.

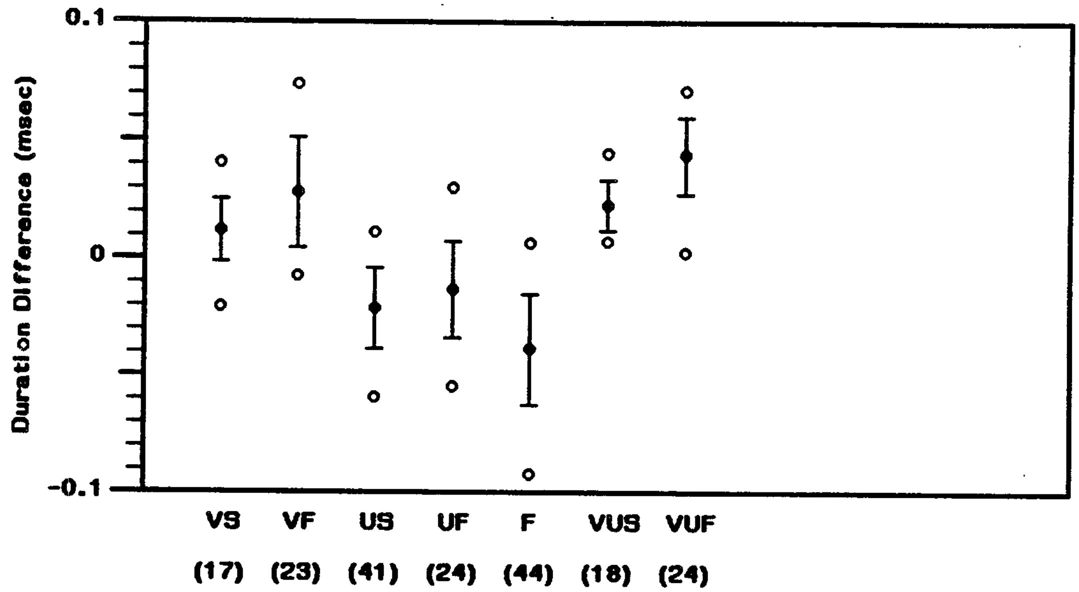


Figure 3.1: Statistical Summary of Minimal Pair Experiments

This display summarizes the results of minimal pair experiments which measured differences in the duration of nasal consonants in two different contexts. From left to right, the contexts are: word final vs. nasal voiced stop (VS), word final vs. nasal voiced fricative (VF), word final vs. nasal unvoiced stop (US), word final vs. nasal unvoiced fricative (UF), word initial vs. unvoiced fricative nasal (F), nasal voiced stop vs. nasal unvoiced stop (VUS), and nasal voiced fricative vs. nasal unvoiced fricative (VUF). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

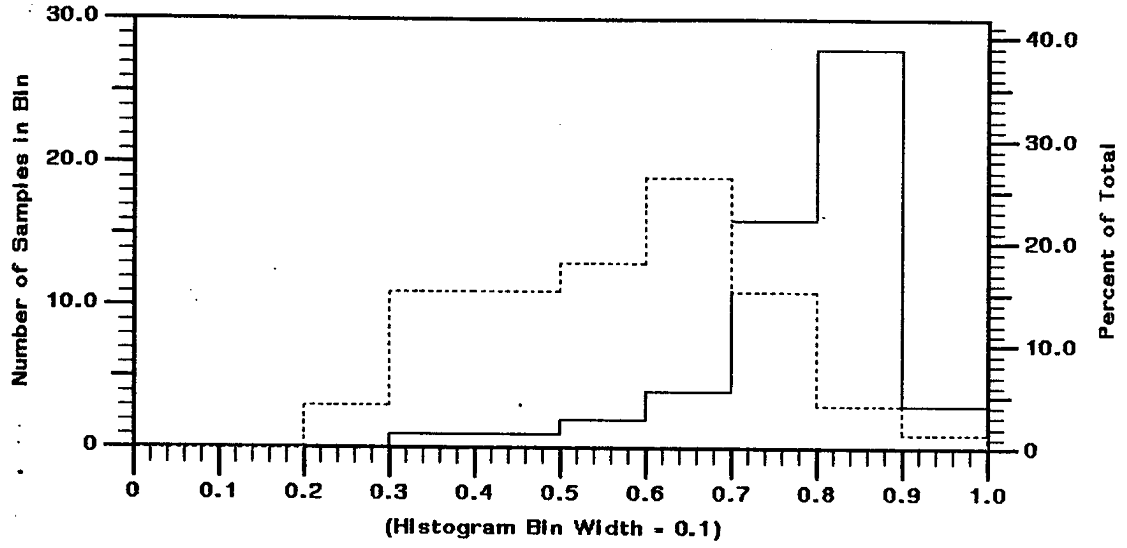


Figure 3.2: Voicing Discrimination in Nasal Stop Consonant Sequences

The solid lines outline the fraction of time that the nasal murmur occupies the total nasal murmur and stop gap duration of voiced stops (55 tokens). The dashed lines outline the same fraction for voiceless stops (72 tokens).

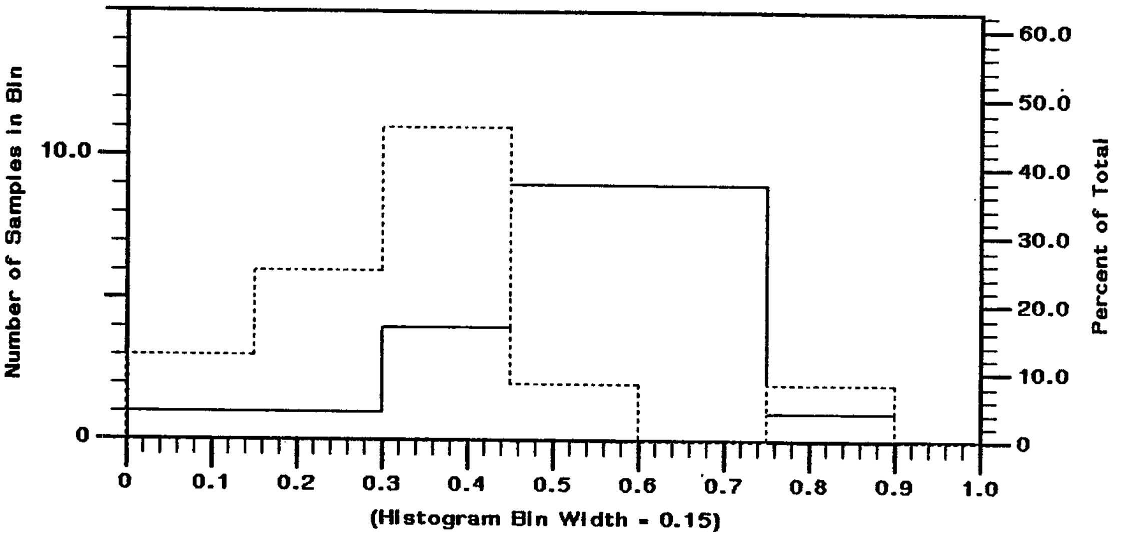


Figure 3.3: Voicing Discrimination in Stop Nasal Consonant Sequences

The solid lines outline the fraction of time that the nasal murmur occupies the total stop gap and nasal murmur duration of voiced stops (25 tokens). The dashed lines outline the same fraction for voiceless stops (24 tokens).

General Results

In an attempt to establish global duration values of different contexts, the database was reduced to a set of monosyllabic words, with the exception of intervocalic nasals. These words were subdivided into broad contexts. It was found that the duration of nasal murmurs produced by male speakers were affected by context slightly more than females. Global duration values for the two groups are shown in figures 3.4 and 3.5. Note that the average durations of female speakers are greater than the male counterparts in every context. Figure 3.6 summarizes the durations of nasal murmurs in a singleton environment, and those in a cluster with another consonant, for all speakers in the database. Nasal murmurs in a singleton environment had an average duration of 65 msec. Nasal murmurs in a cluster with a voiceless consonant had an average duration of 40 msec, while those in a cluster with a voiced consonant had an average duration of 75 msec. Thus, voiceless clusters tend to shorten nasal murmur duration, while voiced clusters tend to lengthen nasal murmur duration.

In general, fricative nasal consonant clusters had the shortest nasal murmur durations in the database. Figure 3.7 illustrates the distributions of the nasal murmur duration of prevocalic nasals in a syllable initial position, and in a fricative cluster. Fricative consonant clusters also exhibited a period of epenthetic silence between the fricative and the nasal murmur, which was due to a mistiming of the articulators. This period of silence is a very robust acoustic cue for detecting the presence of a nasal consonant (it can also be present in fricative glide clusters such as *slack*) when the nasal murmur is very short. The average duration of the period of silence was found to be 30 msec, as indicated in figure 3.8.

Discussion

Previous studies of nasal murmur durations have been primarily concerned with homorganic nasal stop consonant clusters [79], [62]. All of these investigations have shown that the duration of the nasal murmur is substantially longer when preceding a voiced stop than a voiceless stop. Minimal pair differences range from 25 to 70 msec. The average values found in this research are in the lower end of these values. This is probably because many of these studies measured the durations of nasal consonants in stressed, monosyllabic words, sometimes spoken in isolation. Thus, one would expect tokens spliced out of continuous speech to have shorter durations.

Other researchers have noted the differences in duration of the nasal murmur between male and female speakers [77]. The general finding is that when nasals form a cluster with another consonant, the nasal murmur duration of female speakers is not affected to the same degree as those of male speakers.

It should be emphasized that the majority of the duration statistics were gathered on a subset of the database. With the exception of intervocalic nasal consonants, poly-syllabic words were not included. As one might expect, a minimal pair experiment found that the durations of nasal murmurs in a poly-syllable environment, were shorter than those in a mono-syllable environment, as illustrated in figure 3.9. Thus it would be difficult to apply the knowledge of duration of nasal murmurs to the field of speech recognition, unless one was able to obtain details of the particular context of the nasal consonant under consideration. The fact that the rate of speech itself can vary substantially, further limits the usefulness of duration as a speech recognition parameter.

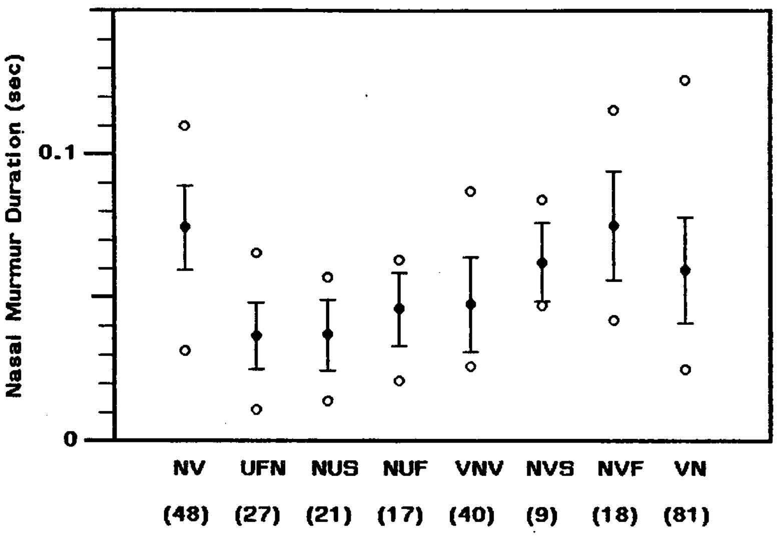


Figure 3.4: A Summary of Nasal Consonant Durations for Male Speakers

This display summarizes nasal murmur durations of male speakers for particular phonetic environments. From left to right they are: singleton prevocalic nasals (NV), fricative nasal clusters (UFN), nasal unvoiced-stop consonant clusters (NUS), nasal unvoiced-fricative clusters (NUF), intervocalic nasals (VNV), nasal voiced-stop consonant clusters (NVS), nasal voiced-fricative clusters (NVF), and singleton post-vocalic nasal consonants (VN). The average value is indicated by a filled circle. Vertical lines indicate one standard deviation. The open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

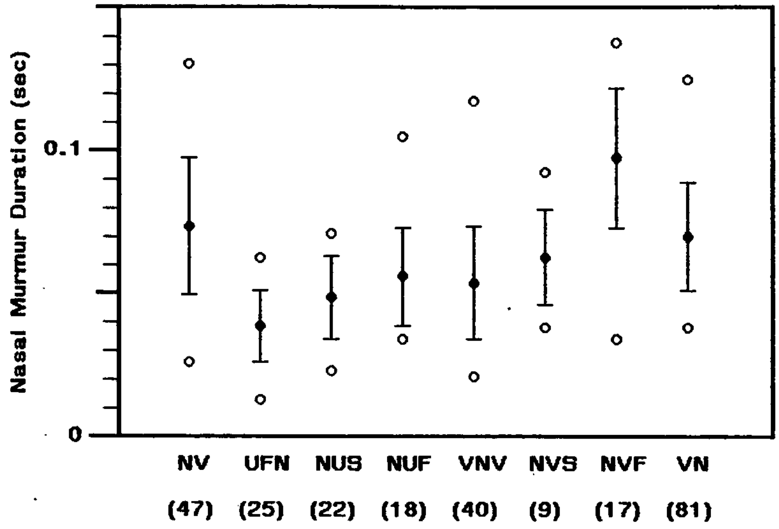


Figure 3.5: A Summary of Nasal Consonant Durations for Female Speakers

This display summarizes nasal murmur durations of female speakers for particular phonetic environments. The contexts are the same as those described in figure 3.4. The average value is indicated by a filled circle. Vertical lines indicate one standard deviation. The open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

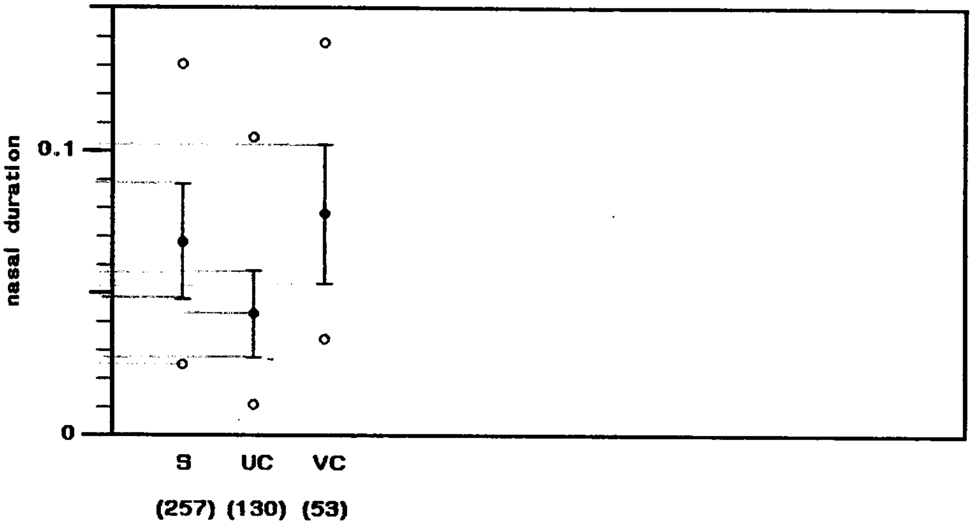


Figure 3.6: A Summary of Nasal Consonant Durations

This display summarizes nasal murmur durations of nasal consonants in a singleton environment (S) as opposed to those in a cluster with a unvoiced consonant (UC) or a voiced consonant (VC). The average value is indicated by a filled circle. Vertical lines indicate one standard deviation. The open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

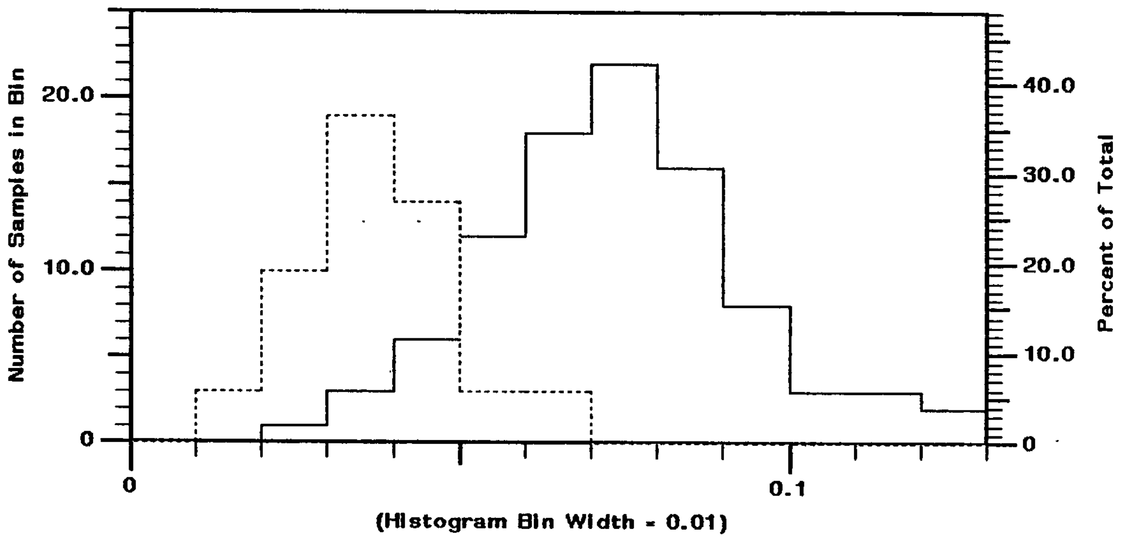


Figure 3.7: Durations of Prevocalic Nasal Consonants

The solid lines outline the duration of nasal consonants in a word initial position (95 tokens). The dashed lines outline the durations of nasal consonants which form a fricative nasal cluster (52 tokens). Thus, this display compares words like *knitt*, and *snit*.

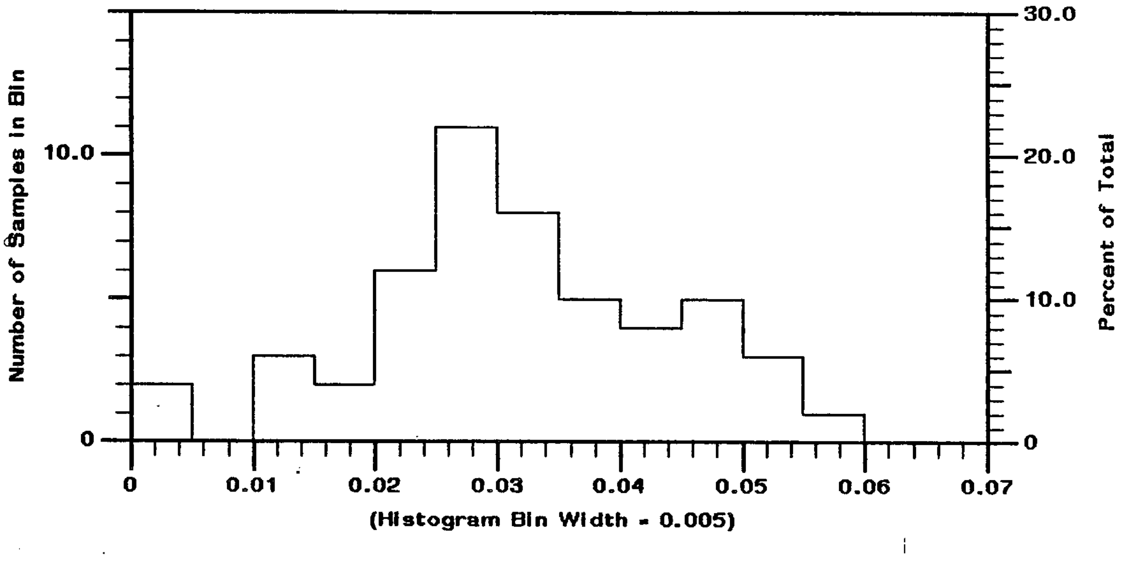


Figure 3.8: Epenthetic Silence Duration of Fricative Nasal Consonant Clusters

The solid lines outline the duration of the period of epenthetic silence between the fricative and the nasal consonant in fricative nasal clusters, as found in the word *snit* (50 tokens).

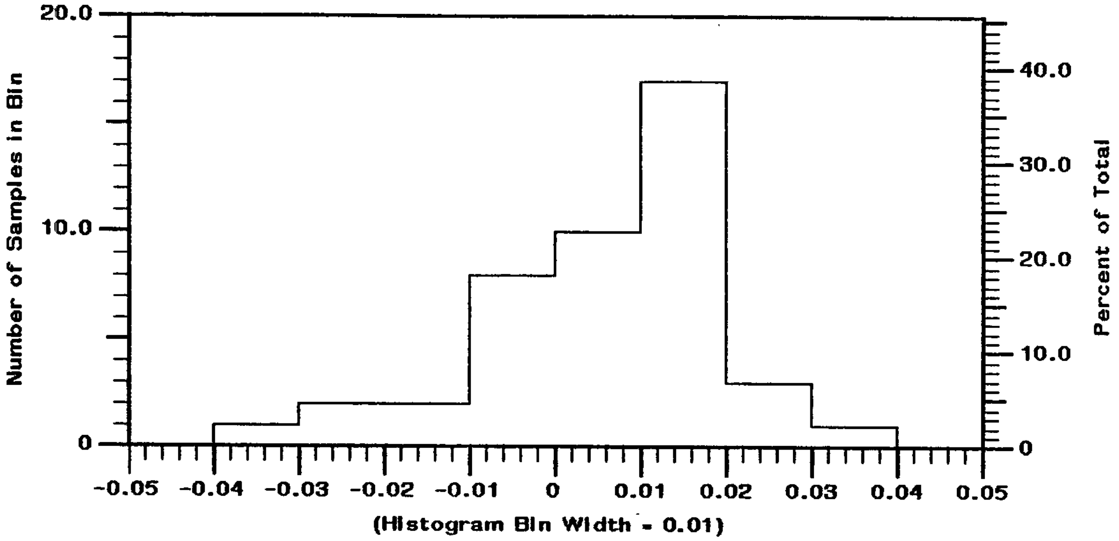


Figure 3.9: Duration Differences of Nasal Consonants between Mono and Poly-Syllabic Words

The solid lines outline minimal pair duration differences between the nasal murmur in a mono-syllabic word and that in a poly-syllabic word, as in the pair *bend/bending* for example (44 tokens).

3.1.2 A Study of Nasal Consonant Energy

The first energy experiment conducted measured an energy difference, calculated by subtracting the average total energy in the nasal murmur from the average total energy in the adjacent sonorant. Figure 3.10 contains a histogram of this energy difference, plotted in dB, for all of the nasal consonants in the database. Since this energy difference is almost always positive, it can be concluded that the nasal murmur is consistently weaker than an adjacent sonorant.

On closer inspection, there are several other observations which may be made about energy differences. Some of these have been illustrated in figure 3.11, which presents a statistical summary of energy differences of nasal consonants in different contexts. As indicated in the figure, there appears to be only minor differences in the energy difference due to vowel quality (front, back, high, or low). The most significant separation appears to be between low back vowels, which have an average energy difference of around 11 dB, and high front vowels, which have an energy difference of 7 dB. This observation is more likely due to the fact that low back vowels have more energy than high front vowels, rather than there being any difference in nasal consonant strengths in these two contexts [12], [23].

As illustrated in figure 3.12, nasal consonants in a medial position between two sonorant regions, have a slightly smaller energy difference, ofaround 6 dB, than nasal consonants in other contexts, typically around 10 dB. This is probably due to the fact that medial nasals have strong energy throughout the murmur, since they are surrounded by two sonorants which have strong energy, and do not taper off as would nasals in other contexts. This observation is reinforced by measurements of the nasal murmur stability, discussed shortly, which indicate that the energy of medial nasals is quite steady.

Figure 3.12 also compares the value of the energy difference of the nasal consonants to similar sounds such as liquids and glides, and voice bars (the common name for the period of closure of voiced stop consonants), which are also adjacent to a sonorant region. Although there is some overlap in the distributions, it is clear that on average, nasal consonants have a greater drop in energy than the liquids or glides, and have a smaller difference than voice bars.

As was mentionned in chapter 2, it was not possible to measure a relative energy difference for all voice bars, since many were not immediately adjacent to a sonorant region, being separated by a stop consonant release. However, it is possible to compare the average energy of these two groups. As shown in figure 3.13, the average energy of isolated voice bars tends to be much weaker than voice bars adjacent to a sonorant, and can be discriminated from most nasal consonants on the basis of energy alone. Figure 3.14 presents a statistical summary of the total energy of nasal consonants and similar sounds.

The next parameter which was observed was the energy stability of the nasal consonant. This was measured by calculating the average value of the first difference of the energy in the middle 50% of the nasal murmur. This measure is proportional to calculating the standard deviation of the energy in the murmur. Figure 3.15 illustrates a histogram of the average difference for all of the nasal consonants in the database. Figure 3.16 presents a statistical summary of the average difference for similar speech sounds.

Discussion

The most important point of the analysis of nasal consonant energy, is that nasal consonants tend to be weaker than adjacent sonorants by an average of 10 dB. This result agrees with previous studies of the nasal consonants [24].

For speech recognition, the energy of the nasal consonant has the potential to be a useful parameter, since nasal consonants tend to be stronger than voice bars, and weaker than semivowels.

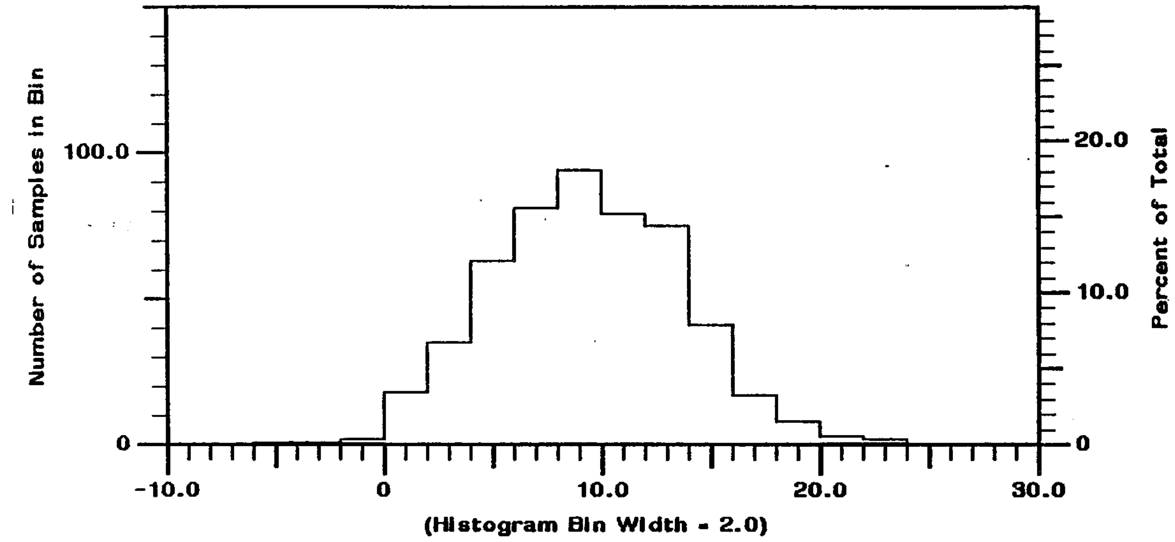


Figure 3.10: Energy Difference of Nasal Consonants

This figure contains a histogram of the energy difference between a nasal consonant, and an adjacent sonorant (520 tokens). Values are plotted in dB.

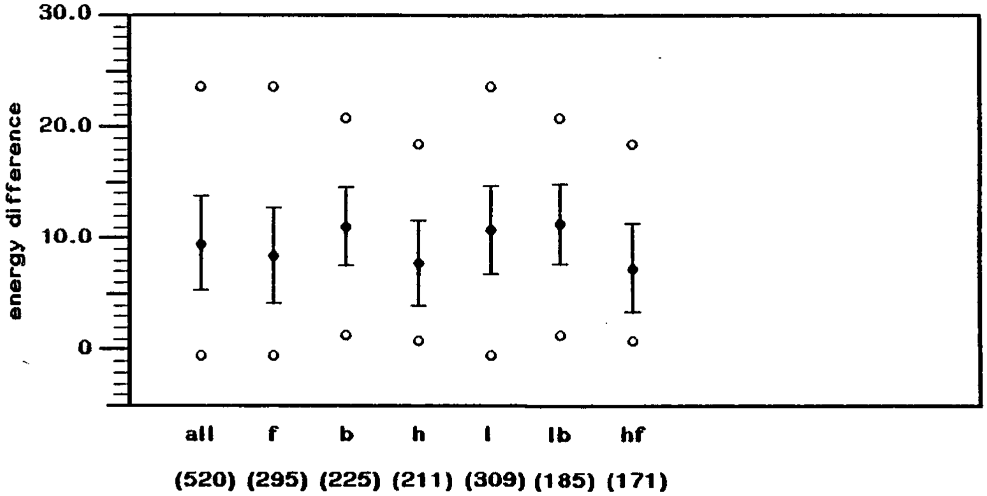


Figure 3.11: Energy Difference Statistics due to Vowel Quality

This display summarizes energy differences of nasal consonants in different contexts. From left to right, they are: all nasal consonants (all), nasals adjacent to front vowels (f), back vowels (b), high vowels (h), low vowels (l), low back vowels (lb), and high front vowels (hf). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display maximum and minimum values. The number of samples in each context are indicated below the display.

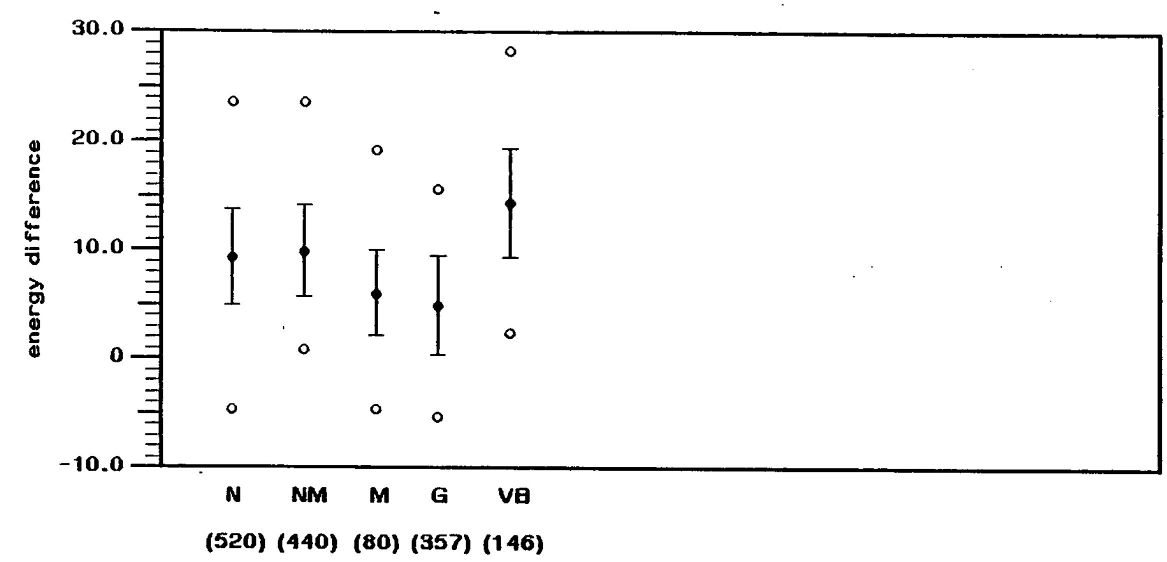


Figure 3.12: Energy Difference Statistics of Similar Sounds

This display summarizes energy differences of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), non-medial nasals (NM), medial nasals (M), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation and the open circles display maximum and minimum values. The number of samples in each context are indicated below the display.

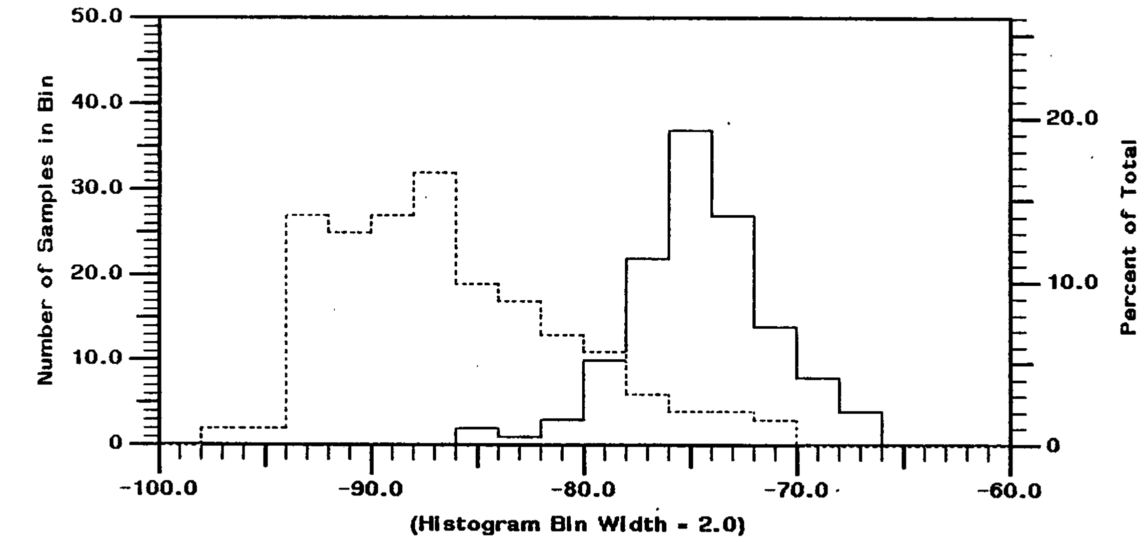


Figure 3.13: Energy of Voice Bars

The solid lines outline average energy for voice bars adjacent to a sonorant (146 tokens). The dashed lines outline the average energy of voice bars not adjacent to a sonorant (192 tokens).

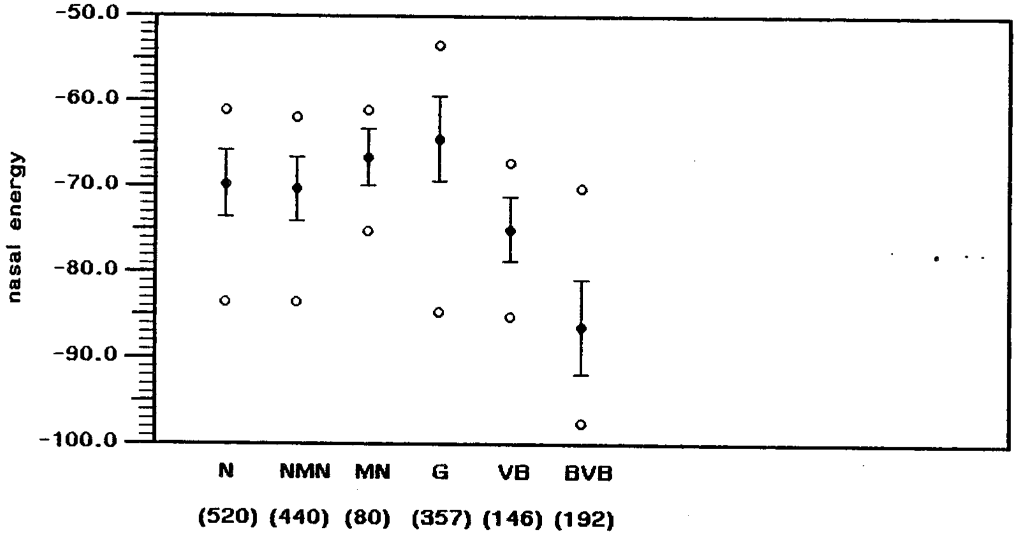


Figure 3.14: Statistics of Energy

This display summarizes energy differences of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), non-medial nasals (NMN), medial nasals (MN), liquids and glides (G), voice bars adjacent to a sonorant (VB), and voice bars not adjacent to a sonorant (BVB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

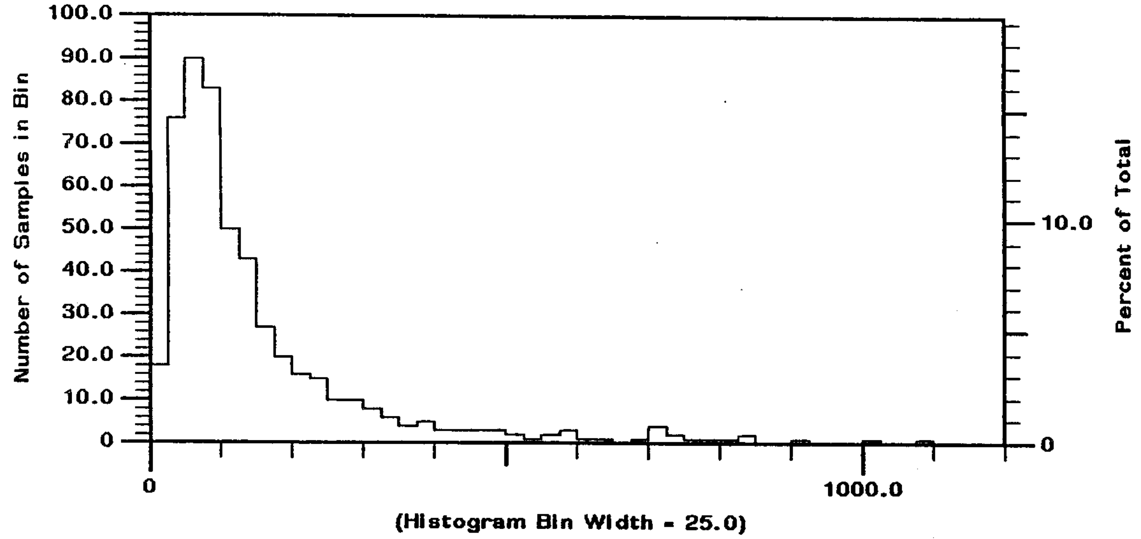


Figure 3.15: Energy Stability of Nasal Consonants

This figure contains a histogram of the average first difference of energy, calculated in the middle region of the nasal murmur (520 samples). Values are plotted in dB per second.

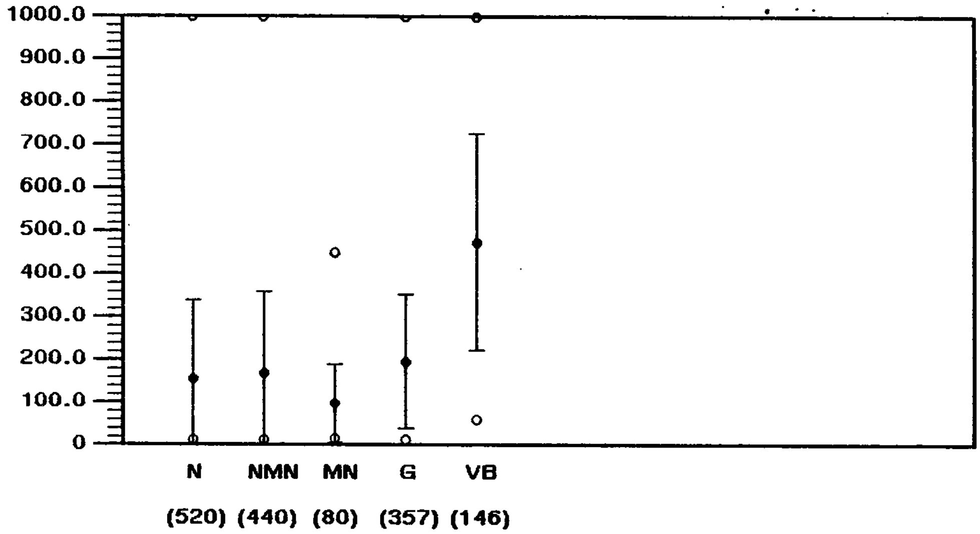


Figure 3.16: Statistics of Energy Stability

This display summarizes the average energy change of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), non-medial nasals (NMN), medial nasals (MN), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

3.1.3 A Study of Nasal Consonant Spectra

Once the spectral shape of the nasal murmurs were normalized with respect to total energy, it was possible to measure an average spectral shape using the techniques described in chapter 2. Figures 3.17 and 3.18 show average spectra of the three nasal consonants for one speaker. In general, the spectral shapes of the nasal consonants were found to be highly speaker dependent. This is not surprising, since the size of the nasal and sinus cavities can vary greatly from speaker to speaker. Although subtle differences could be detected between the three nasal consonants for any given speaker, all three nasal consonants tended to have similar spectral shapes, as indicated by these figures. This observation is in agreement with that made by Fujimura, who also found little differences among the magnitude spectra of the three nasal consonants [15].

In general, nasal consonant spectra were characterized by a low frequency energy which dominated the spectrum. Several measures were made in order to quantify this property.

Figure 3.19 plots the frequency of the largest peak in the spectrum for all of the nasal consonants in the database. As may be seen, the low frequency energy was not only nearly always the largest in the spectrum, it was nearly always centered between 200 and 350 Hz as well. Figure 3.20 displays the results of a measurement which calculated the percentage of the time that a nasal consonant had a resonance centered between 200 and 350 Hz (all values are scaled by 100). The majority of nasal consonants had percentage values close to 1.0. Figure 3.20 also plots this percentage for semivowels as well. Figure 3.21 presents a statistical summary of the percentage values for nasals, semivowels, and voice bars. From these figures, it may be concluded that the presence of a low frequency resonance is a necessary, but not sufficient, condition for the identification of a nasal consonant. In other words, if a token does not have a value near 1.0 for the calculation, it is extremely unlikely that it is a nasal consonant.

As previously mentioned, this low resonance energy dominates the overall spectrum of the nasal murmur. Figure 3.22 displays the results of a measure which calculated the relative amount of energy in the low frequency region of the spectrum (below 500 Hz) for all of the nasal consonants, and semivowels. This measure may also be obtained by plotting the normalized amplitude of the low resonance directly. Clearly, the majority of the energy in the nasal consonant is found in the low frequency region. Figure 3.23 presents a statistical summary of this measure for nasal consonants, semivowels, and voice bars.

The final characteristic of the low resonance which was quantified, was an abrupt decrease in energy in the frequencies immediately above the low resonance. Figure 3.24 displays the results of a measure which calculated the amount of low frequency energy (below 350 Hz) relative to local adjacent energy (350 to 1000 Hz). This measure was not overly sensitive to the actual locations of the frequency boundaries. This measure could also be obtained by spectral weighting functions, such as center of gravity measures in the low frequency region. Figure 3.25 presents a statistical summary of this measure for nasal consonants, semivowels, and voice bars. From this figure, it is apparent that semivowels have less of a drop than nasal consonants, and voice bars have slightly more. In fact, this measure is very effective in seperating nasal consonants from most semivowels.

Finally, a measure of the spectral stability of the nasal consonant spectra was also made. As was indicated by figure 2.5, the spectra of nasal consonants were found to be quite stable at frequencies below 1000 Hz. There are several ways that this can be measured, including measuring the standard deviation from a spectral average, or a spectral weighting function, such as the center of mass. Figure 3.26 displays a histogram of the average deviation of the normalized low frequency energy (below 1000 Hz). The distribution of voice bars is also displayed for comparison. Figure 3.27 presents a statistical summary of this measurement for similar sounds.

Discussion

The analysis of the nasal consonant spectra primarily verified the results of previous studies, which indicated that the spectrum is dominated by a low frequency energy around 300 Hz.

There were several properties of nasal consonants which were difficult to quantify successfully. For instance, it is commonly known that the nasal consonant has several higher frequency resonances, and that the resonance bandwidths are generally higher than in vowel-like sounds. Furthur, nasal consonants have an antiformant, whose frequency location depends on the place of articulation. The problem with attempting to measure any of these parameters is that resonances do not always show up as peaks in the magnitude spectrum, and antiformants will not necessarily show up as valleys in the spectrum. This phenomenon results from pole zero cancellation, as Fujimura illustrated.

For speech recognition purposes, the most robust spectral property of the nasal consonant would appear to be a steady low frequency resonance, which is centered between 200 and 350 Hz. The most useful characteristics of this resonance are the percentage and height measures, since they are able to discriminate nasal consonants from other sounds with similar acoustic properties. The measure of low resonance amplitude is more useful at discriminating between nasal consonants and sounds which do not have a predominance of low frequency energy.

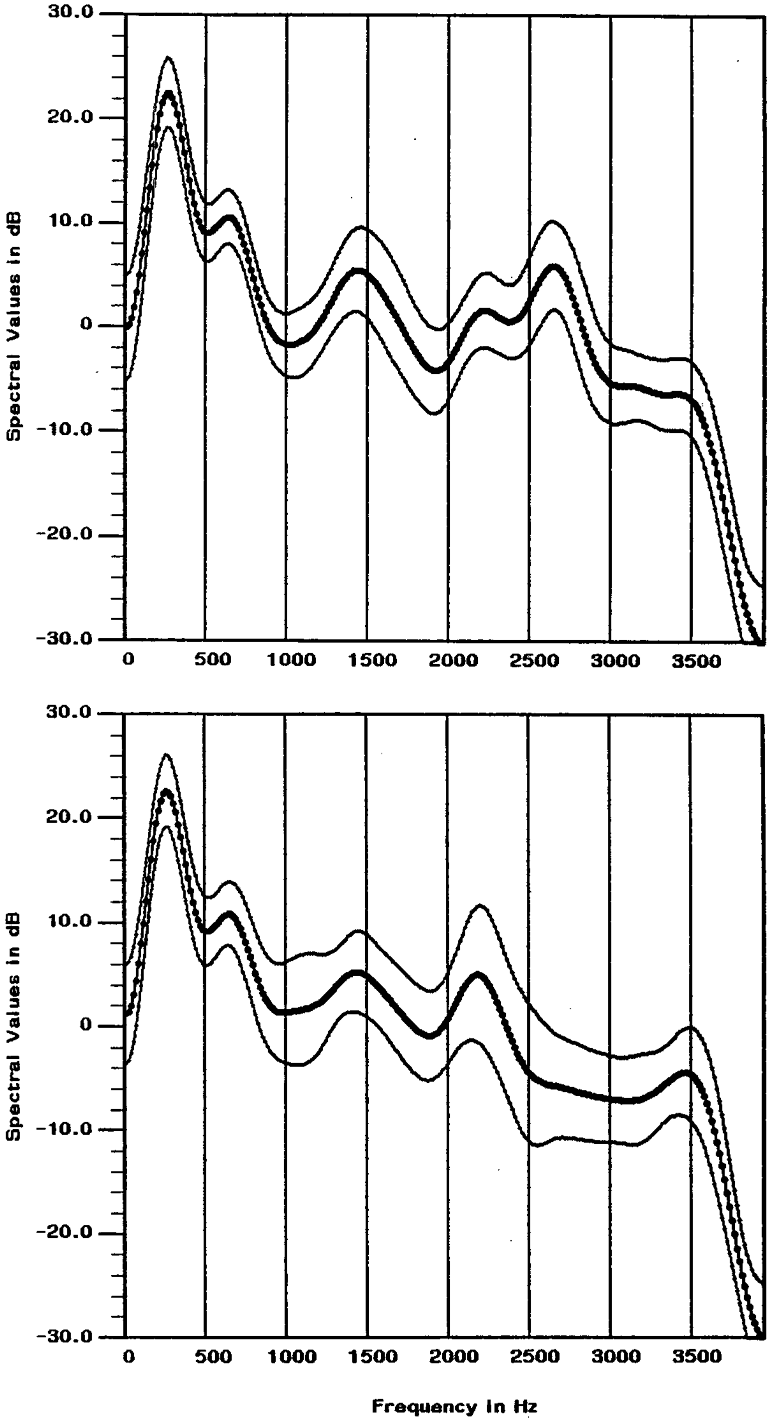


Figure 3.17: Average Spectral Shape of /n/, and /m/

This top display presents a statistical summary of the normalized smoothed spectra of the nasal consonant /n/, for a male speaker. The bottom display presents a summary of an /m/. spoken by the same speaker. The average spectral shape, shown by the dark line, is surrounded by lines which represent one standard deviation from the mean.

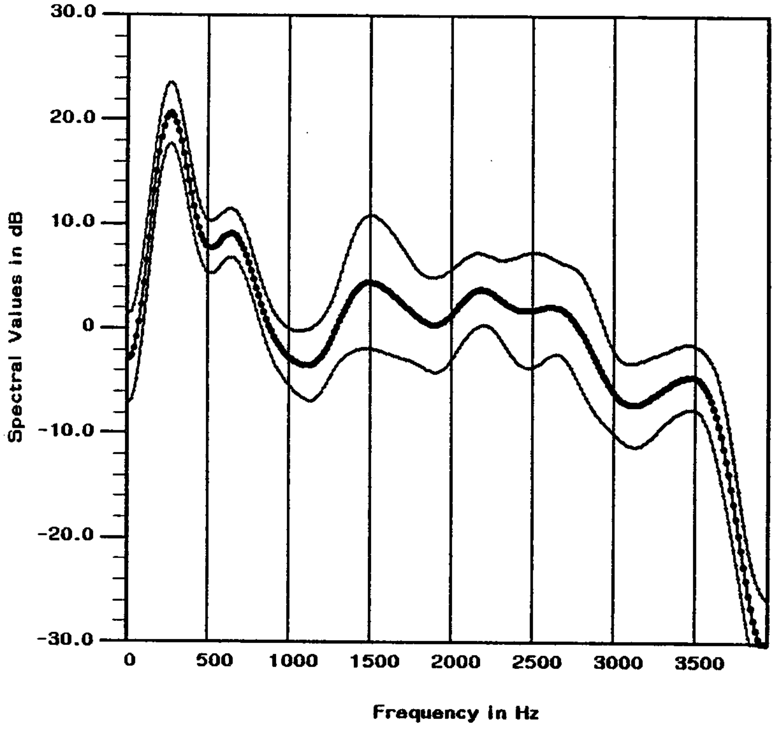


Figure 3.18: Average Spectral Shape of /ŋ/

This display presents a statistical summary of the normalized smoothed spectra of the nasal consonant /ŋ/, for a male speaker. The average spectral shape, shown by the dark line, is surrounded by lines which represent one standard deviation from the mean.

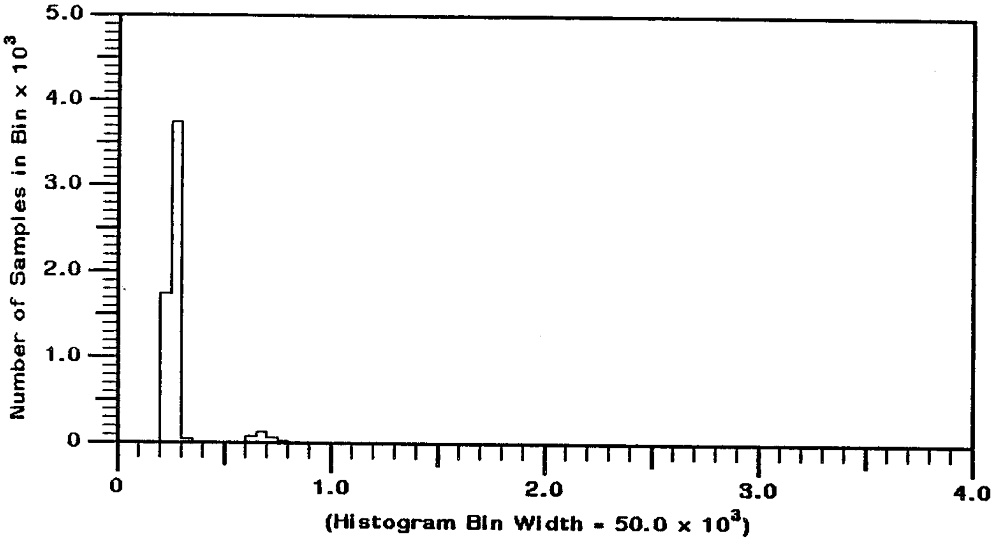


Figure 3.19: Frequency of Largest Spectral Peak in the Nasal Consonant

This display contains a histogram of the frequency of the largest spectral peak in the nasal consonant (6092 tokens). Values were collected for multiple spectra from each nasal consonant, and are plotted in thousands of Hz.

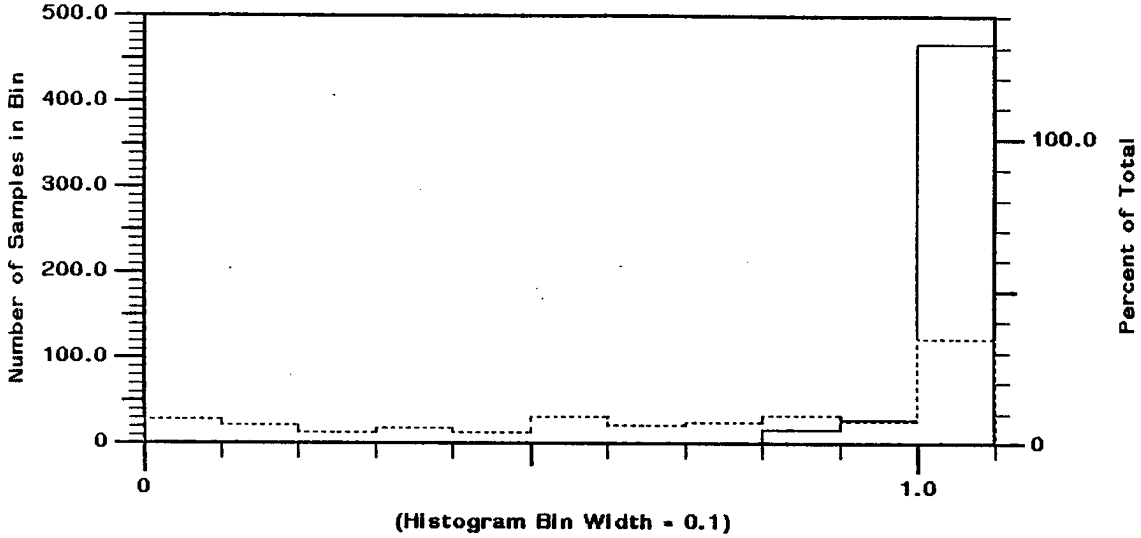


Figure 3.20: Low Resonance Percentage

This display contains a histogram of the percentage of time in a nasal consonant that there was a low frequency resonance centered in between 200 and 350 Hz. The solid lines are the distributions of nasal consonants (520 tokens). The dashed lines are the distributions of semivowels (357 tokens). Values between 1.0 and 1.1 have a value of 1.0

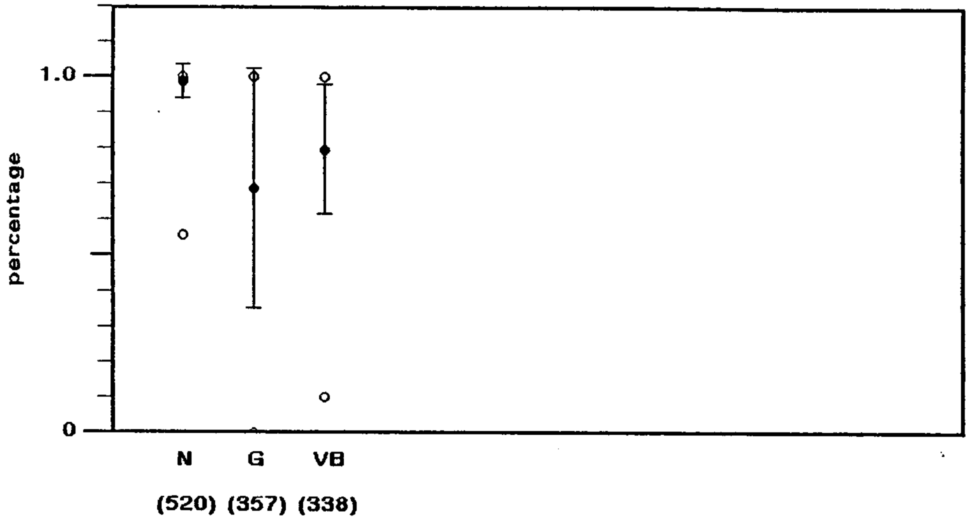


Figure 3.21: Statistics of Low Resonance Percentage

This display summarizes the low resonance percentage of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

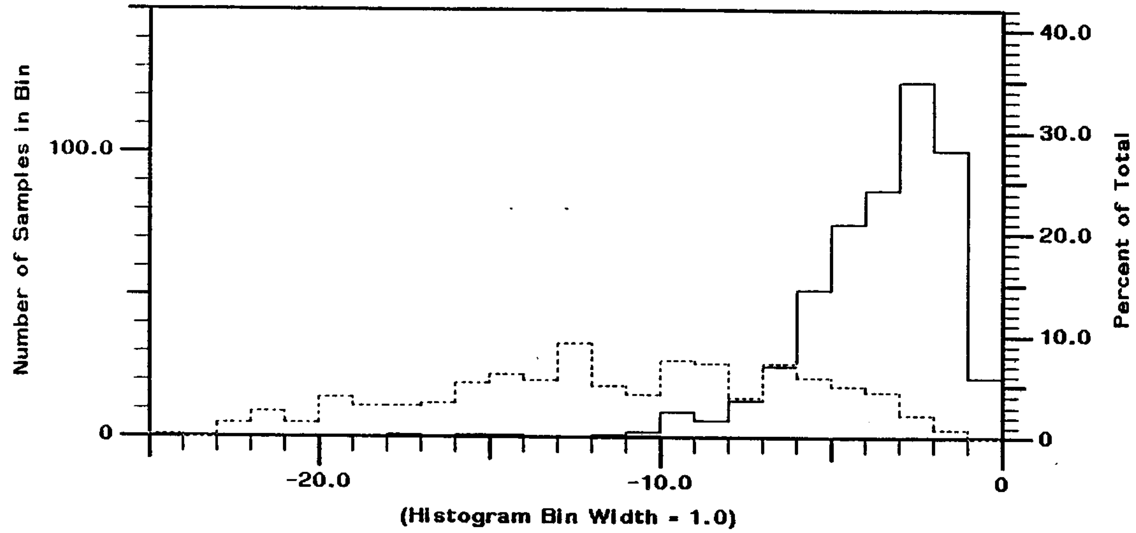


Figure 3.22: Low Resonance Amplitude

This display contains a histogram of the relative amplitude of low frequency energy. The solid lines are the distributions of the nasal consonants (520 tokens). The dashed lines are the distributions of semivowels (357 tokens). Values are plotted in dB.

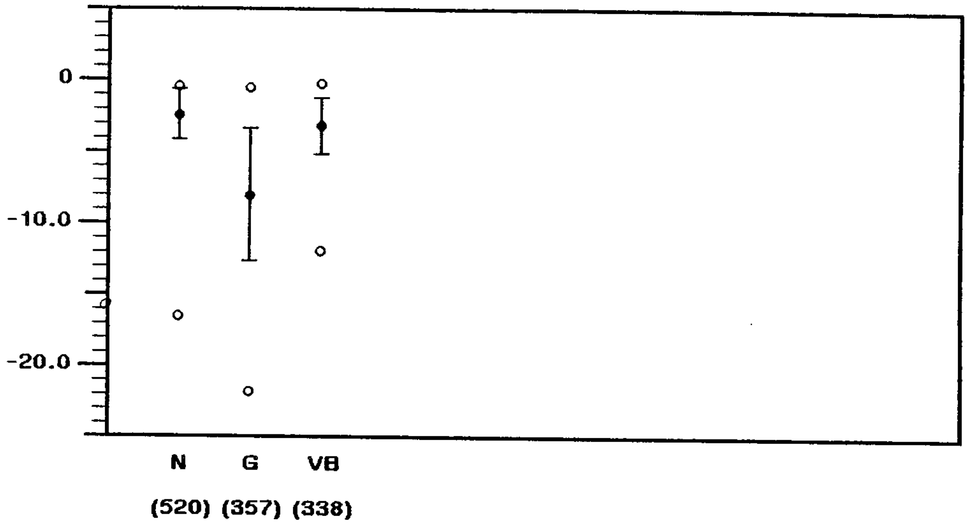


Figure 3.23: Statistics of Low Resonance Amplitude

This display summarizes the low resonance amplitude of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

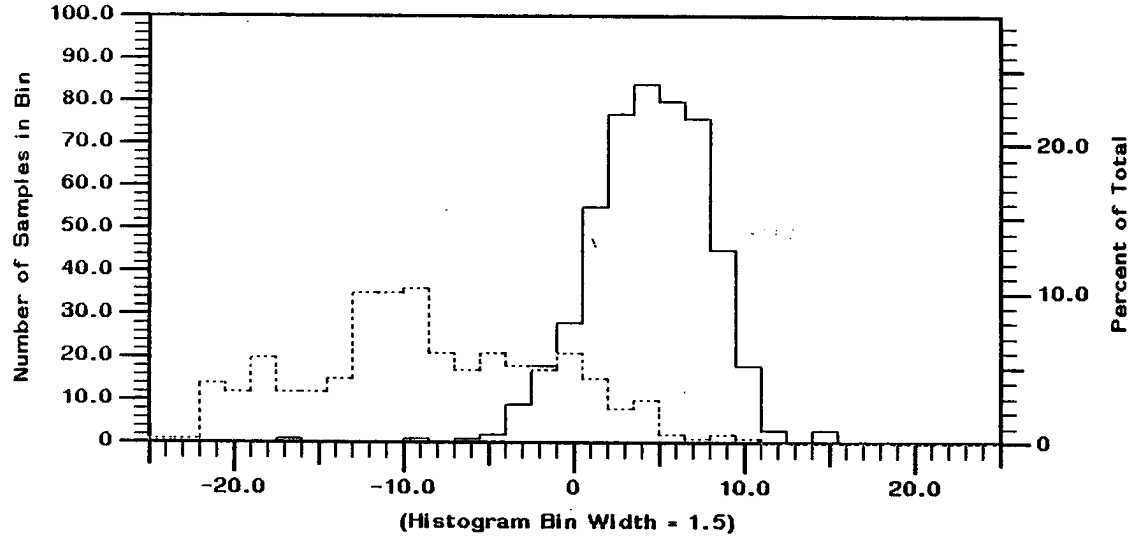


Figure 3.24: Low Resonance Height

This display contains a histogram of the local relative amplitude of the low frequency resonance. The solid lines are the distributions of the nasal consonants (520 tokens). The dashed lines are the distributions of semivowels (357 tokens).

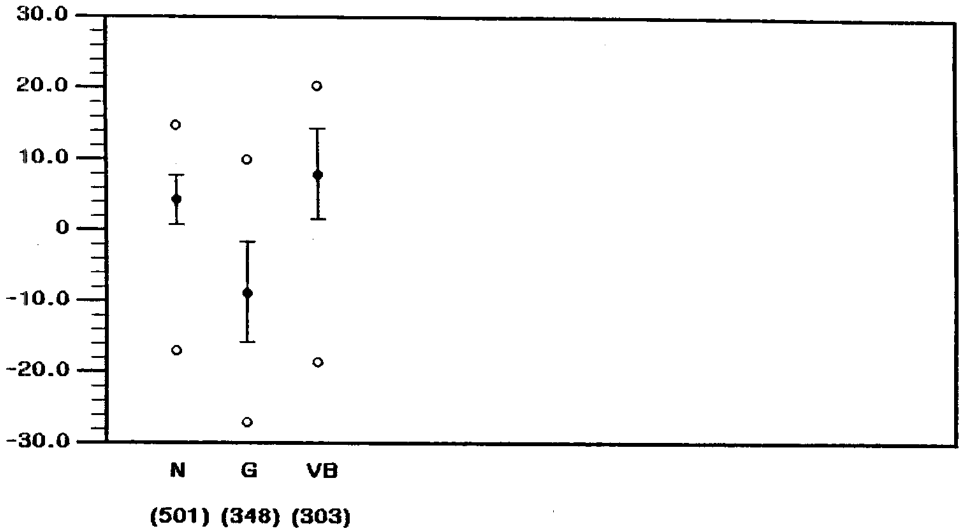


Figure 3.25: Statistics of Low Resonance Height

This display summarizes the low resonance height of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

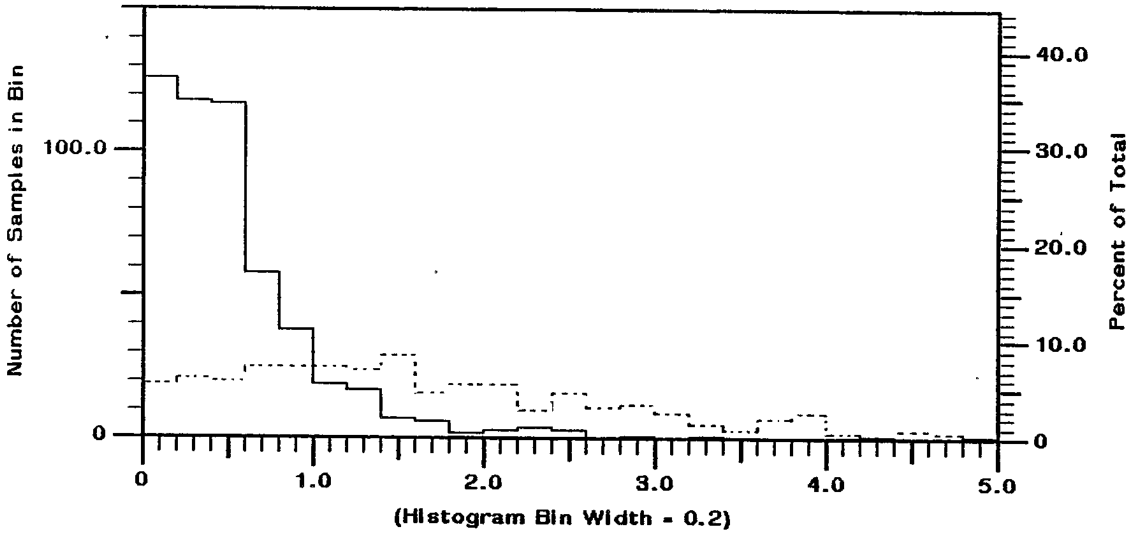


Figure 3.26: Spectral Stability

This display contains a histogram of the standard deviation of relative amplitude of low frequency energy. The solid lines are the distributions of the nasal consonants (520 tokens). The dashed lines are the distributions of semivowels (357 tokens). Values are plotted in dB.

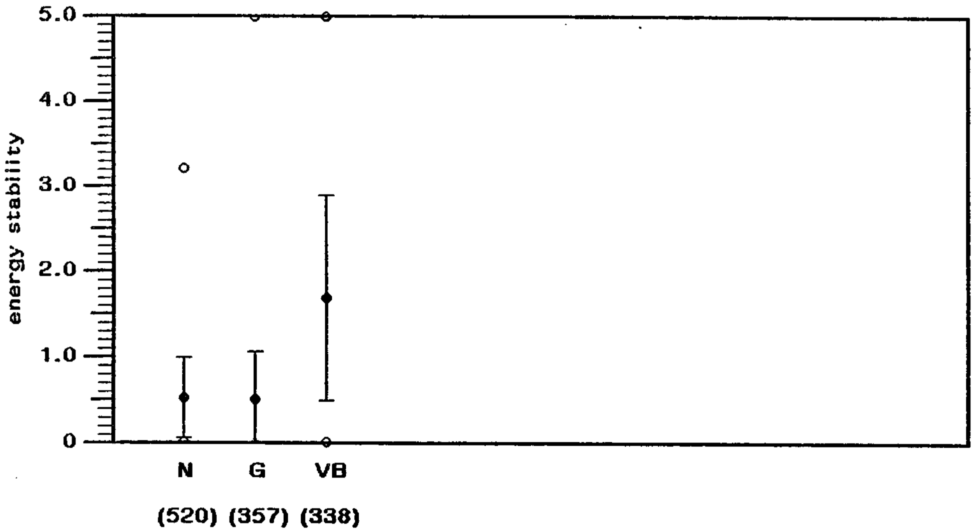


Figure 3.27: Statistics of Spectral Stability

This display summarizes the spectral stability of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display

3.2 Analysis of Nasalized Vowels

A study of nasalized vowels is more complicated than a study of nasal consonants. Since nasalized vowels are not distinguished phonemically from oral vowels in American English, it is perfectly legitimate to nasalize a vowel in any phonetic context. It is therefore not possible to separate nasalized and oral vowels by a phonetic transcription alone. Nasalization could be established by measuring the airflow from the nasal cavities. Vowels with any nasal coupling would then be easily separated from completely oral vowels. Subsequently, an acoustic study of the speech waveforms could establish properties which separate these two groups of vowels.

However, the goal of this research is to establish properties of nasalized vowels which may be used to help detect the presence of a nasal consonant. Thus, it is more useful to classify vowels using the criterion of whether or not they are adjacent to a nasal consonant. The goal is then to establish acoustic differences between these two groups of vowels, making the acoustic study one of *relative* nasalization. The underlying assumption is that when vowels are next to a nasal consonant, they are nasalized *more* than they would be otherwise. In this research then, nasalized vowels are defined as those vowels adjacent to a nasal consonant, while non-nasalized vowels are those vowels that are not adjacent to a nasal consonant.

Although the results of such a study are potentially beneficial to speech recognition, there are several complicating factors. First, in American English, a vowel is often nasalized whenever a nasal consonant is present somewhere in the syllable nucleus, even if it is not immediately adjacent to the vowel. For instance, the /I/ in the word *film* will tend to be nasalized. By the definition used in this research, /I/ would be classified as a non-nasalized vowel. By its context however, it is likely to be nasalized. Since the nature of these vowels is somewhat ambiguous, they were filtered out of the database in order to reduce the amount of noise they might cause in the measurement distributions. This excluded about 200 vowels from the acoustic analysis. Another alternative would have been to classify them as nasalized vowels, since it is likely that they were indeed nasalized.

Although filtering operations will reduce the number of nasalized vowels in a non-nasal context, it will never eliminate all such cases, as is illustrated for the word *back*, shown in figure 3.28. This is because some speakers tend to naturally nasalize all vowels, and also because low vowels are quite often slightly nasalized, independent of context. Clearly, the challenge of the acoustic study is to establish measures which can automatically differentiate between the /æ/ in *back*, and the /æ/ in a word like *mack*, also shown in figure 3.28.

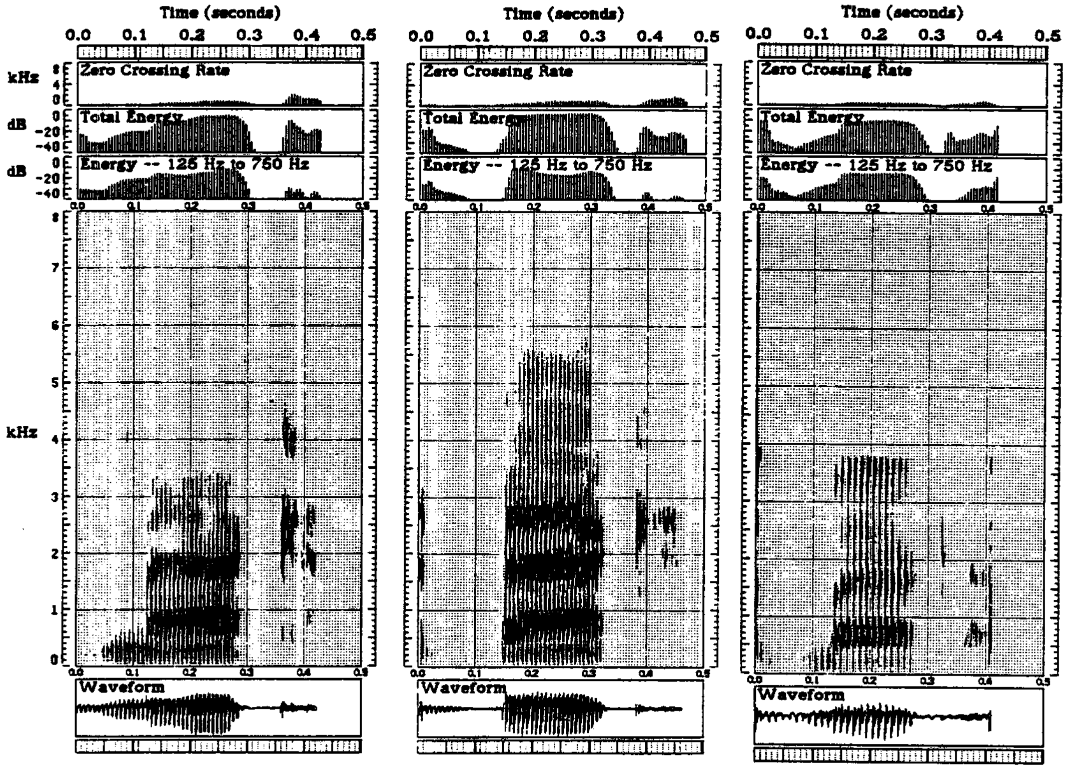


Figure 3.28: Spectrograms of the words *mack*, and *back*

Left: A spectrogram of the word *mack* spoken by a male speaker. Middle: A spectrogram of the word *back*, spoken by the same speaker. Right: A spectrogram of the word *mack*, spoken by a different male speaker.

Another difficulty with a study of nasalized vowels is that different speakers nasalize to various degrees. Thus, one persons nasalized vowel could have the same characteristics as another's non-nasalized vowel. This phenomenon, also illustrated in figure 3.28, smears measurement distributions, and illustrates the difficulties associated with speaker-independent nasalized vowel identification.

For the acoustic analysis, there are a few procedures used to reduce the magnitude of this problem. First, the initial analysis is conducted on a speaker by speaker basis, to eliminate speaker-independent complications. In fact, in order to eliminate as much speaker and context variability as possible, the initial portion of the analysis is restricted to observing relative differences between minimal word pairs such as *skip* and *skimp*. This allows the observation of acoustic differences which are introduced by the presence of the nasal consonant.

A perceptual study, performed on a subset of the database, is also used to aid the analysis. Given a vowel token spliced out of the speech waveform, subjects must decide whether or not the vowel is next to a nasal consonant.[1] Once each vowel is given a nasality rating, acoustic characteristics may be correlated to establish a perceptual credibility. Of course, there are several issues which need to be considered in such a test, including whether or not untrained subjects know what a nasalized vowel is, or how natural the tokens are. However, the scores were found useful for guiding the initial study.

The inherent dynamic quality of nasalized vowel spectra also complicate the acoustic analysis. Unlike nasal consonants, nasalized vowels are not necessarily steady state sounds due to either the nature of nasalization (lowering or raising of the velum), or of the vowel itself (dipthongs). In either of these cases, the net effect is that the acoustic characteristics change with time. Averaging procedures, used throughout the analysis of the nasal consonants, are not adequate in this case. Figure 3.29 shows a spectrogram of the word *made*, where the low frequency regions of the vowel are clearly changing with time.

[1] The perceptual study of the database is reported in detail in the next chapter.

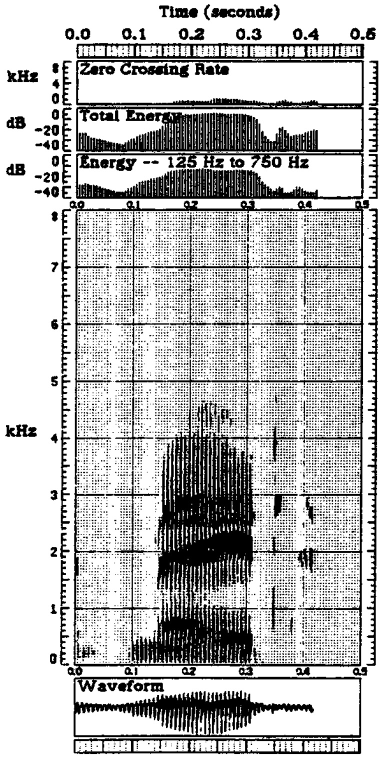


Figure 3.29: A Spectrogram of the word *made*

Of course, it is possible to try and track useful characteristics of the vowel (such as the resonance frequencies) for the duration of the vowel. This method was not used because such systems tend to be rather fragile, especially in nasalized vowels. Instead, the vowel was divided into *subsegments*, so that averaging procedures could be used in each subsegment to reduce measurement noise, yet changes between the different subsegments of the vowel, caused by increasing nasalization for example, would still be measureable. After some experimentation it was decided to use three subsegments in each vowel. Thus, whenever a measurement of some parameter was made on a vowel, there were three values returned. Each value represented an average of the parameter in one of the three, equally spaced, vowel subsegments.

The following sections report results of the study of the durational, and spectral characteristics of nasalized vowels.

3.2.1 A Study of Nasalized Vowel Duration

Since there are many contextual factors which can influence the duration of vowels, it would be unreasonable to expect to be able to distinguish nasalized vowels from non-nasalized vowels on the basis of duration alone. However, a minimal pair experiment was performed to establish if indeed there were any differences in duration. For this experiment, vowels in a nasal consonant context, such as *meat*, were paired with vowels in either a stop or fricative consonant context, such as *beat*. The difference measure was calculated by subtracting the two vowel durations.

Figure 3.30 displays a histogram of the difference in duration for all of the vowel pairs. On average, vowels appear to be shortened by approximately 10 msec when they are put into a nasal consonant context. The spread of this distribution weakens the strength of this statement however. On closer inspection of the data, it appears that the greatest difference is between vowels in a fricative nasal cluster, such as *smack* versus *sack*, where the average difference is nearly 20 msec.

When the nasal consonant formed a post-vocalic cluster with a stop, or fricative consonant, the vowel duration was observed to vary with the voicing of the clustering consonant. Statistics for minimal pair duration differences between words such as *bend* and *bent*, or *ones* and *once*, may be found in figure 3.31. Vowels in a nasal stop consonant cluster, were observed to be lengthened by 30 msec on average, when the stop consonant was voiced. Vowels in a nasal fricative consonant cluster, were observed to be lengthened by 10 msec on average, when the fricative consonant was voiced. Note that this durational change is much less significant than that observed in the nasal consonants in the same circumstances.

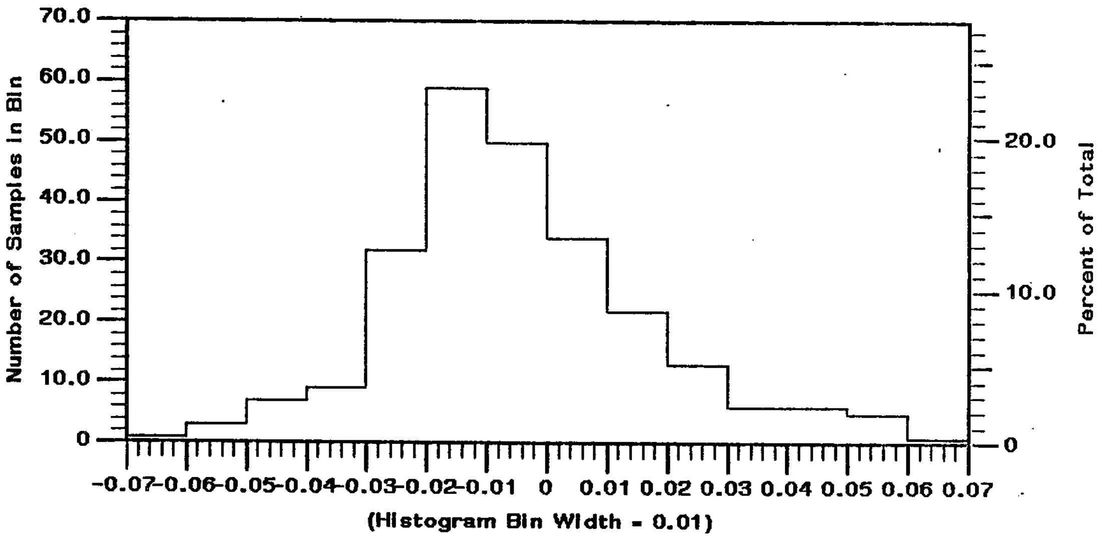


Figure 3.30: Vowel Duration Differences

The solid lines outline minimal pair duration differences between vowels in a nasal consonant context, such as the word *bent*, and those in a stop or fricative consonant context, such as the word *bet* (253 tokens).

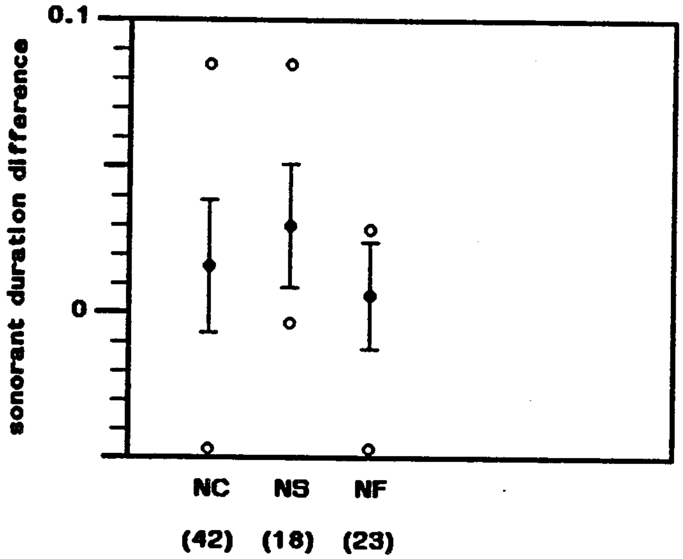


Figure 3.31: Vowel Duration Differences due to Voicing

This display summarizes the difference in vowel duration of minimal pairs in different voicing contexts. From left to right they are: all nasal consonant clusters (NC), nasal stop clusters (NS), nasal fricative clusters (NF). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

3.2.2 A Study of Nasalized Vowel Spectra

The spectral analysis of nasalized vowels was carried out in a manner similar to that of the nasal consonants. In the first stage of analysis, the goal was to establish differences between nasalized and non-nasalized vowels by comparing some form of average spectra. On the basis of these observations, general discriminating properties could be proposed and quantified using utterances of the database. The following sections describe the sequence of steps followed for the spectral analysis.

Spectral Averaging

Using the multiple spectra averaging technique described for the analysis of nasal consonants, statistics were collected for nasalized and non-nasalized vowels of each speaker. Initially, two average spectra were computed for each vowel (nasal context and non-nasal context). Figure 3.32 shows average spectra for an /æ/ for a male speaker.

Although there was a danger of smearing a significant amount of information by the averaging procedure, these plots were quite informative. The most noticeable difference between the nasalized and non-nasalized vowels was in the low frequency regions of the magnitude spectrum . On average, it was found that non-nasalized vowels had one resonance in the first formant region, while nasalized vowels had two. Of the two resonances found in the nasalized vowel, one could always be associated with a first formant. This resonance was labelled the "first resonance". The extra resonance, which could appear above, or below the first resonance, depending on the vowel height, was labelled the "nasal resonance", although it was clear that this resonance was not always a result of nasal coupling. In figure 3.33 for instance, which contains a nasalized, and non-nasalized /æ/ from the words *camp*, and *cap*, the first resonance is located at about 700 Hz, for both the nasalized and non-nasalized vowels. The nasal resonance is located near 250 Hz for both vowels as well. In figure 3.34 however, which contains a nasalized, and

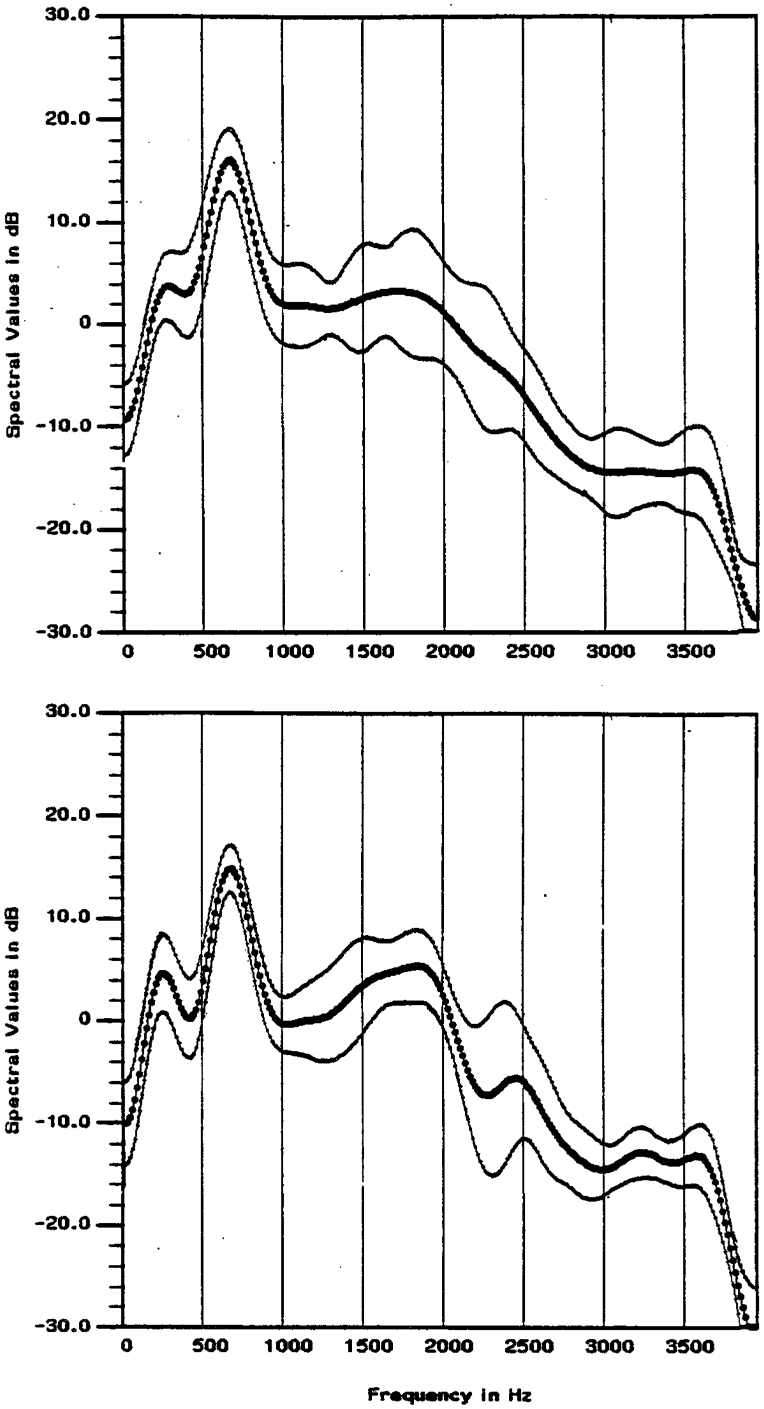


Figure 3.32: Average Spectral Shape of /æ/

The top display presents a statistical summary of the normalized, smoothed spectra of the non-nasalized /æ/ of a male speaker. The bottom display presents a summary of the nasalized /æ/ of the same speaker. The average spectral shape, shown by the dark line, is surrounded by lines which represent one standard deviation from the mean.

a non-nasalized /i/ from the words *technique*, and *beat*, the situation is different. Here the first resonance is located at about 350 Hz for both non-nasalized and nasalized vowels, and the extra resonance is located at 700 Hz.

In general it was observed that the nasal resonance would appear above the first resonance only for very high vowels, where the first resonance was centered below 400 Hz. Otherwise, the nasal resonance appeared below the first resonance.

Unfortunately, many non-nasalized vowels were observed to have a nasal resonance, as is clear from these figures.[2] This means that it is not always possible to distinguish nasalized from non-nasalized vowels by measuring the fraction of time that there is a nasal resonance in the vowel.

Fortunately, it was found that the nasal resonance was noticably more "distinct" in a nasalized vowel. There were two ways in which "distinctness" was manifested in the spectrum. First, the magnitude of the nasal resonance could increase relative to the first resonance. This could be caused by the first resonance decreasing in amplitude, or the nasal resonance increasing, or both. Second, the dip between the nasal resonance and the first resonance could deepen. Thus, if a non-nasalized and a nasalized vowel both happened to have an extra resonance, it is possible to discriminate between them by measuring the relative strength of the nasal resonance to the first formant. The previous two figures both provide good examples of how the nasal resonance is more distinct in the nasalized vowel.

Another observed characteristic of nasality was a smearing of the first resonance itself. In fact, when an extra resonance was not present, as was occasionaly observed in a nasalized vowel, a measure of the spread of energy about the first resonance was found to be the best indication of nasalization.

In summary then, by observation of spectra, a set of qualitative characteristics of vowel nasalization was proposed. Due to the variability of the environment, none

[2] The vowels produced by female speakers tended to have a low resonance in any context. This property was due to breathiness more than nasalization.

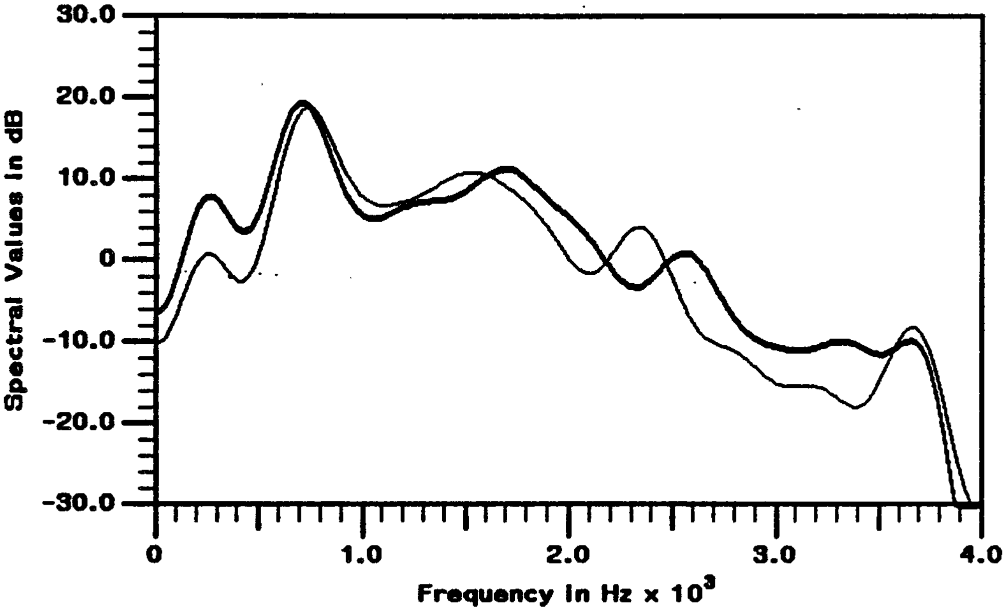


Figure 3.33: Overlay of Nasalized and Non-nasalized /æ/

This display contains spectra of the vowel /æ/ taken from the words *cap*, and *camp*. The light line is for the vowel in the non-nasalized context.

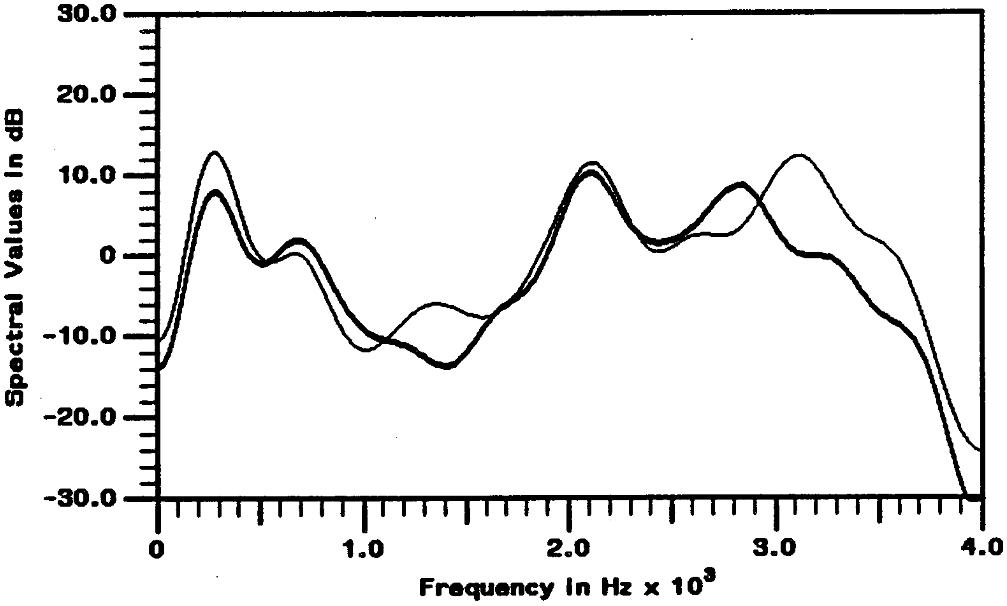


Figure 3.34: Overlay of Nasalized and Non-nasalized /i/

This display contains spectra of the vowel /i/ taken from the words *beat*, and *technique*. The light line is for the vowel in the non-nasalized context.

of these characteristics was present in a nasalized vowel at all times. However, taken in combination, these properties were able to discriminate between nasalized and non-nasalized vowels. The next step was to quantify these observations. A set of algorithms was developed which were able to automatically extract these measures of nasalization. The details of the algorithms may be found in Appendix D. The following sections present the results of the quantitative analysis of nasalized vowels.

Minimal Pair Experiments

As a first step at quantifying the qualitative descriptions of nasalized vowels, the differences between the vowels of minimal word pairs such as *ben* and *bed*, *mack* and *back*, were observed. This procedure effectively eliminated speaker-dependent and vowel-dependent variability. The results of these minimal pair experiments have been summarized below.

The first parameter measured was a center of mass of the spectrum below 1000 Hz. A scatter plot of the average value of the center of mass in the middle portion of the vowel is shown in figure 3.35. The horizontal coordinate of a vowel is its value in a nasal environment, such as the /æ/ in *camp*. The vertical coordinate of the vowel is its pair value in a non-nasal environment. In this case the pair word is *cap*, spoken by the same speaker. Any vowel which has the exact same value of center of mass for both contexts will be located on the solid line. If a vowel has a lower center of mass value when it is nasalized, then it will lie above this line. If it has a higher center of mass when it is nasalized, the vowel appears below the line.

In general, it is clear that low vowels such as /æ/ tend to have a lower center of mass when they are nasalized, while high vowels such as /i/ tend to have very little change or a slight increase in center of mass. The change in values are a result in the increase in strength of the extra resonance produced through nasal coupling.

The next measure observed was the standard deviation of the local energy (within 500 Hz) around the center of mass. It was hypothesized that the low frequency energy of nasalized vowels is spread out, due to a weakening of the main formant and a strengthening of the extra resonance. Thus it would be expected that nasalized vowels would have a higher standard deviation than their non-nasalized counterparts. Figure 3.36 shows that in general this is true. The main exception to the rule is the vowel /æ/. This resulted from an artifact of the standard deviation computation which varied slightly with the center of mass value. The fact that /æ/'s have a much lower center of mass value when they are nasalized is enough to lower the deviation values slightly. Since the center of mass is relatively unchanged for other vowels, the standard deviation measure was not influenced to the same degree.

After observing general statistical properties of the low frequency spectra, measurements were made on the actual resonances. From observations of the average spectra, it was clear that nasalized vowels tended to have an extra resonance in the first formant region. Thus the first calculation measured the percentage of the time that there was an extra resonance in the vowel region. A similar calculation looked at the percentages in the three vowel subsegments. This measure was found to be more effective, since it allowed local areas of nasalization to stand out more than would be the case for an overall average. Figure 3.37 shows the value of the maximum percentage of the three subsegment (scaled by 100). In general, it may be seen that nasalized vowels nearly always have a greater maximum percent than their non-nasalized pairs. In fact, most of the nasalized vowels have a value greater than 0.8. Another distribution compares the values of the *minimum* percentage value of the three vowel subsegments, as shown in figure 3.38. Note that the only vowel which still has a high percentage is /æ/, indicating that this vowel nearly always has a low resonance.

Another observed quality of nasalization is the resonance dip, a measure of the drop in energy in between the two resonances, indicating the prominence of the weakest peak. Figure 3.39 plots the maximum value of this dip in the three subsegments of the vowel. Clearly nasalized vowels tend to have a larger dip than their counterparts. This observation strengthens the argument that the extra resonance becomes more distinct as nasalization increases.

The final measure observed compared the relative difference in amplitude between the two resonances as shown in figure 3.40, which plotted the minimum value of the difference for the three subsegments in the vowel. The resonance difference was calculated by subtracting the amplitude of the low resonance from the amplitude of the higher resonance. There are two points to note here. In low vowels, the extra resonance appears below the first formant. Thus the difference value will tend to be positive. As the vowel becomes more nasalized the extra resonance becomes stronger so the difference becomes smaller. In some cases, the extra resonance becomes so large that the difference becomes negative. The exact opposite is true of high vowels when the extra resonance appears above the first formant. In this case the difference starts off negative and, as the extra resonance grows in magnitude, becomes more positive. In the extreme case (never observed), this resonance would be larger than the first resonance, making the difference positive. Thus, the effect of nasalization on the resonance difference depends on the vowel height.

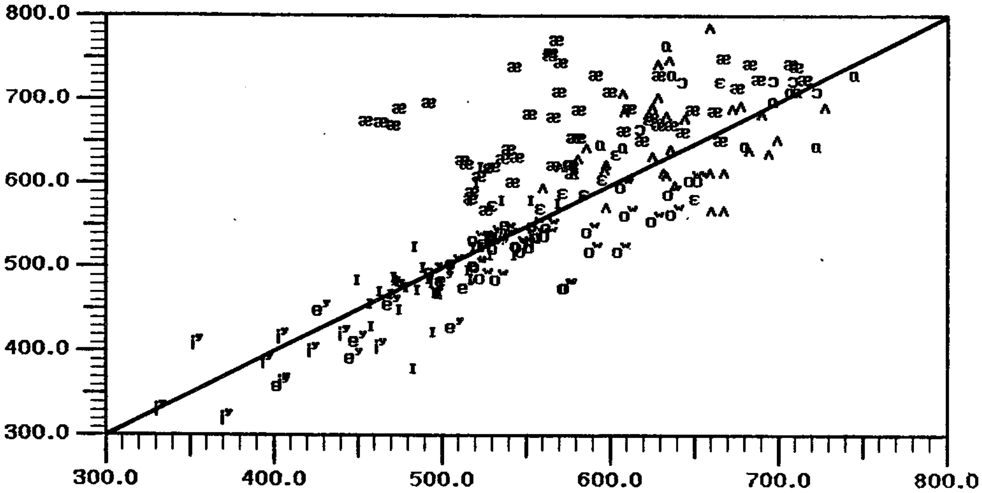


Figure 3.35: Scatter Plot of Center of Mass

This display indicates relative differences in center of mass between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context (such as the /ɛ/ in *bent*). The vertical coordinate of the vowel is its value in a similar, but non-nasal, context (such as the /ɛ/ in *bet*).

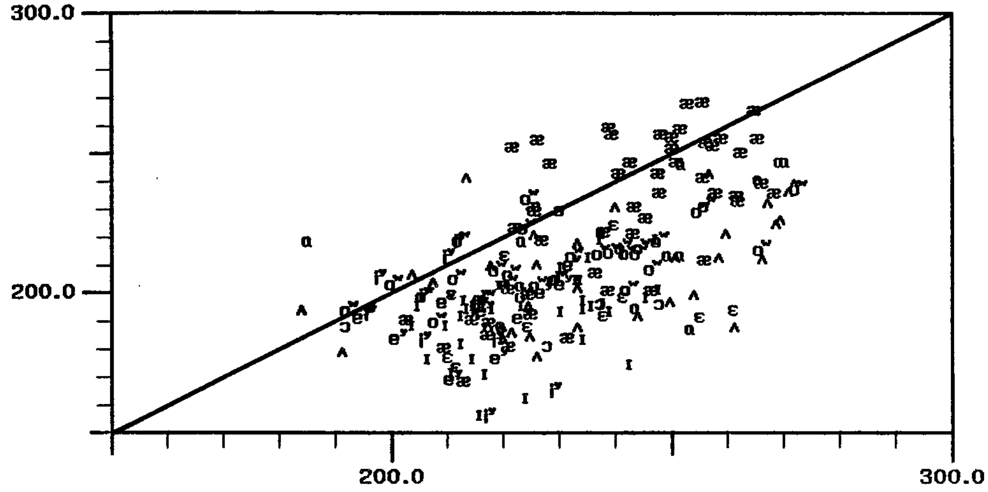


Figure 3.36: Scatter Plot of Standard Deviation

This display indicates relative differences in standard deviation between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

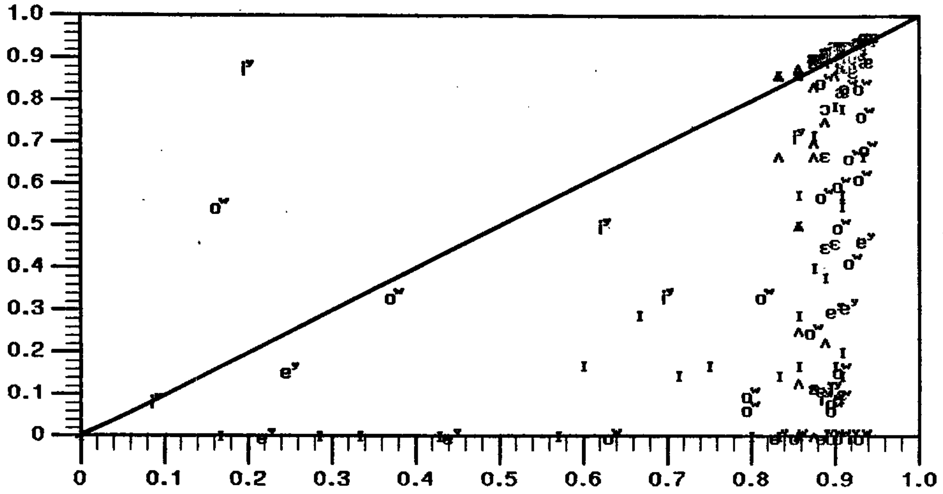


Figure 3.37: Scatter Plot of Maximum Percent

This display indicates relative differences in maximum percentage between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

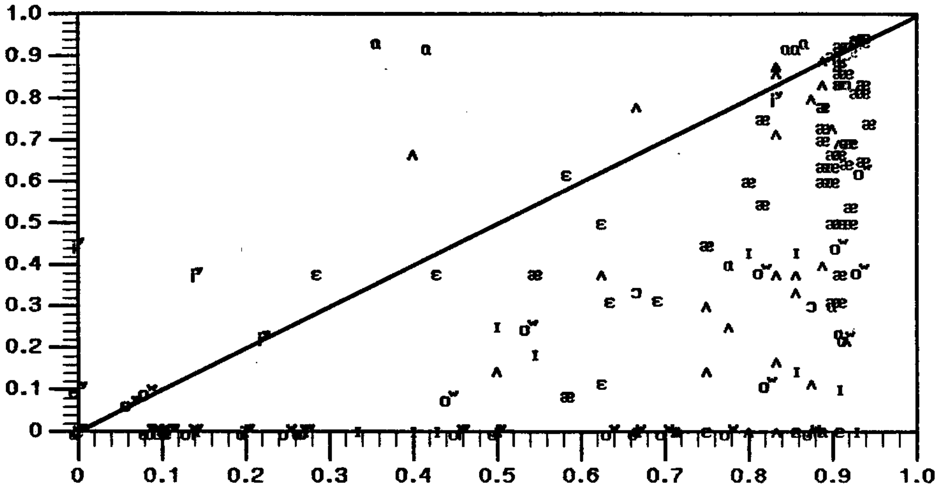


Figure 3.38: Scatter Plot of Minimum Percent

This display indicates relative differences in minimum percentage between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

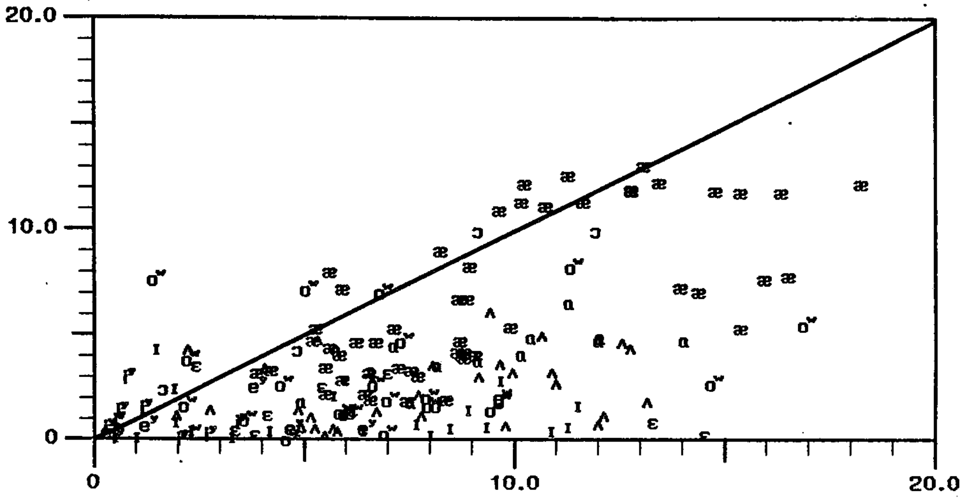


Figure 3.39: Scatter Plot of Maximum Resonance Dip

This display indicates relative differences in resonance dip between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

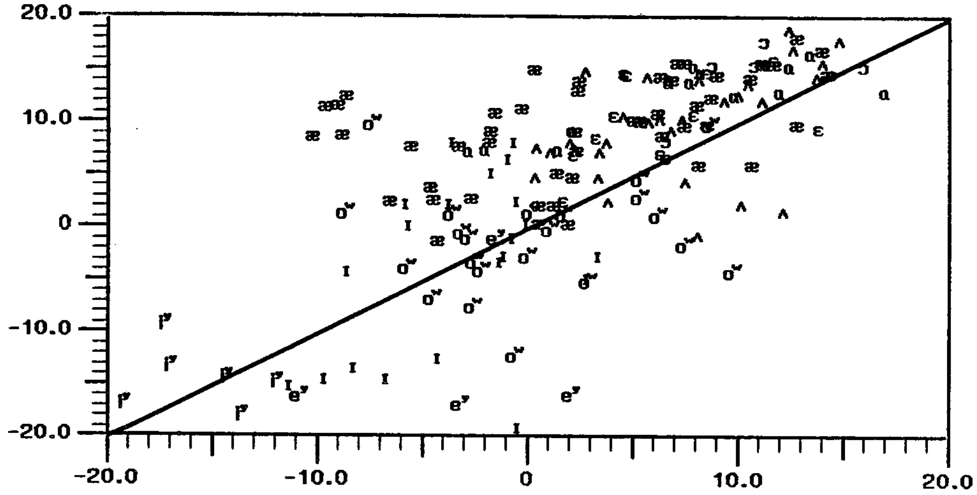


Figure 3.40: Scatter Plot of Minimum Resonance Difference

This display indicates relative differences in resonance difference between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

General Results

Once minimal pair experiments had established some relative results, distributions were made for all of the vowels. It was found useful to retain a high-low distinction in the distributions however, since low vowels tended to have a more distinct nasal resonance than did high vowels. The results of these experiments have been summarized in the following paragraphs.

With the exception of the vowel /æ/, center of mass was not at all effective in discriminating nasalized from non-nasalized vowels.[3] Figure 3.41 shows that center of mass is effective in separating high vowels from low vowels, so this parameter could be of use if vowel height was unknown.

As may be seen in figure 3.42, standard deviation was quite effective in distinguishing nasalized from non-nasalized vowels. In general, nasalized vowels have a higher standard deviation value than non-nasalized vowels. Among each vowel type (high or low), standard deviation does quite well at separating the two groups. Figure 3.43 illustrates the distributions for high vowels.

Figures 3.44 and 3.45 show that the extra resonance percentage measure is also effective in separating nasalized and non-nasalized vowels. From the statistics of the maximum percent region, it is clear that nasalized vowels will always have a high percent value (especially low vowels), while many non-nasalized vowels will not. The minimum percent region shows that low vowels have a resonance throughout the vowel. This is not the case for high vowels, which have a smaller minimum percent. However, since non-nasalized high vowels have even smaller values, this calculation is a good discrimination measure.

Figure 3.46 displays the statistics of the resonance dip measure. Although this calculation is clearly useful, it points out the necessity of being able to

[3] In fact. the change in height of a nasalized /æ/ may be influenced more by phonological rules of American English than by acoustic changes due to nasal coupling [32].

differentiate between high vowels and low vowels, since non-nasalized low vowels have a very similar distribution to nasalized high vowels.

The statistical distributions of the measure of difference, shown in figure 3.47, are perhaps the most difficult to interpret since they appear to overlap. The idea behind this measure was that as the extra resonance became stronger, the difference between it and the first resonance would get smaller. Thus we would expect that as a vowel becomes more nasalized the resonance difference will go to zero. This was certainly true for the low vowels. Unfortunately, there was a problem with computing this for high vowels, since for high back vowels such as /u/, or /o/, the second formant could get confused with a possible low resonance. This resulted in the distribution being rather spread out for non-nasalized high vowels, since there were some very negative values and other very positive values.

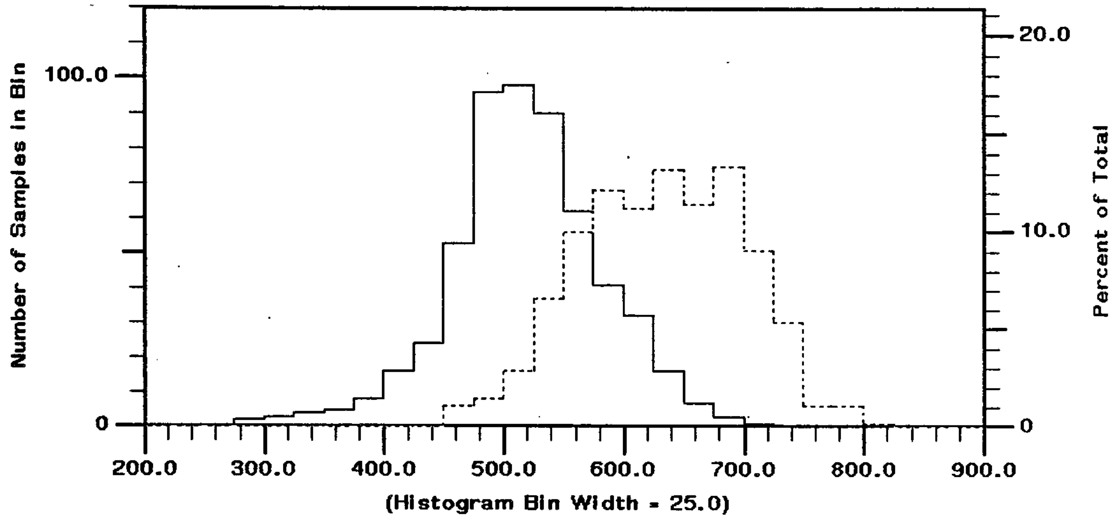


Figure 3.41: Histogram of Center of Mass

This display contains a histogram of the center of mass of all vowels. The dark lines are the distributions of the high vowels (561 tokens). The dashed lines are the distributions of low vowels (561 tokens). Values are in Hz.

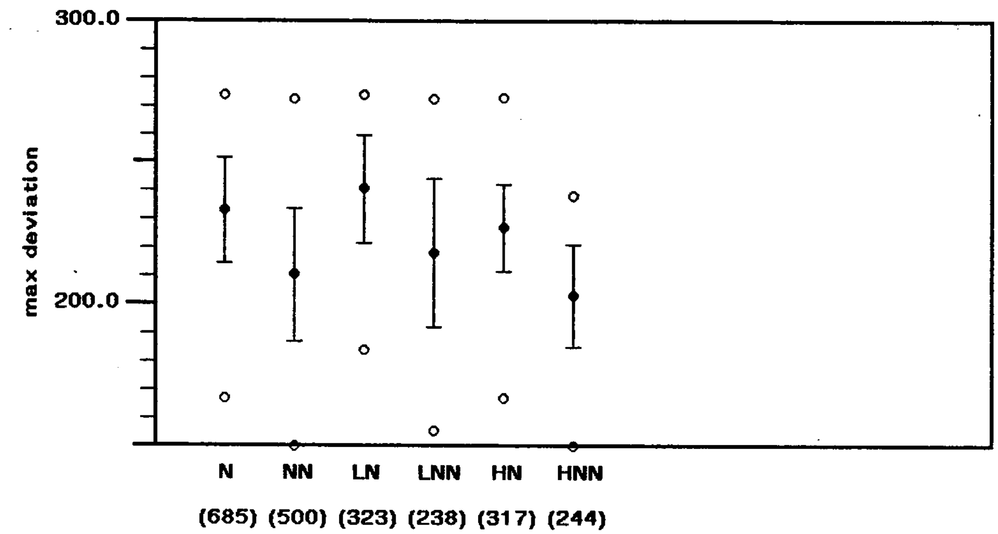


Figure 3.42: Statistics of Maximum Standard Deviation

This display summarizes the standard deviation of nasalized and non-nasalized vowels in different contexts. From left to right they are: all nasalized vowels (N), all non-nasalized vowels (NN), nasalized low vowels (LN), non-nasalized low vowels (LNN), nasalized high vowels (HN), and non-nasalized high vowels (HNN). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

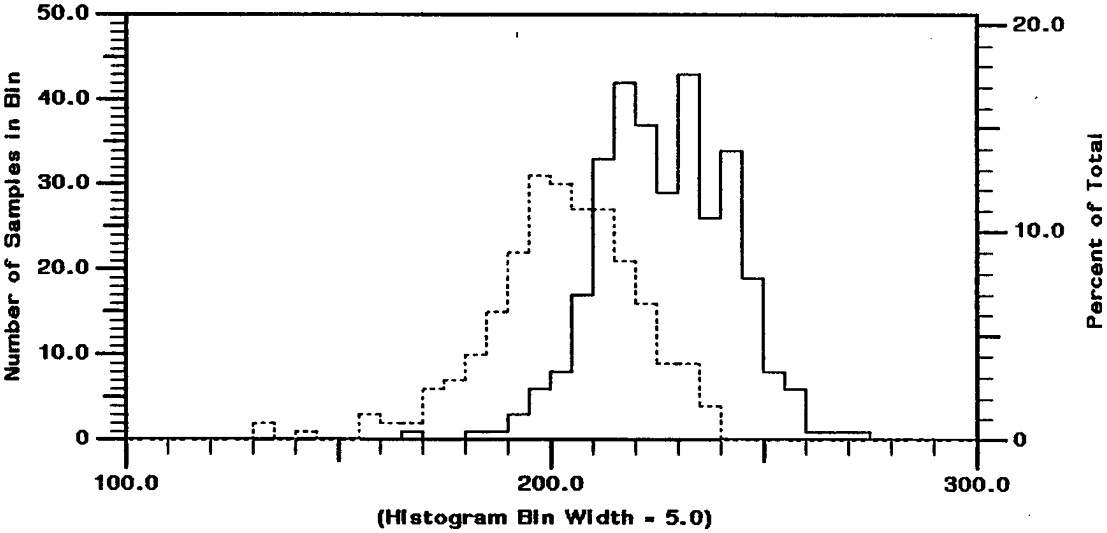


Figure 3.43: Histogram of Standard Deviation of High Vowels

This display contains a histogram of the standard deviation of all high vowels. The dark lines are the distributions of nasalized vowels (317 samples). The dashed lines are the distributions of non-nasalized vowels (244 samples). Values are in Hz.

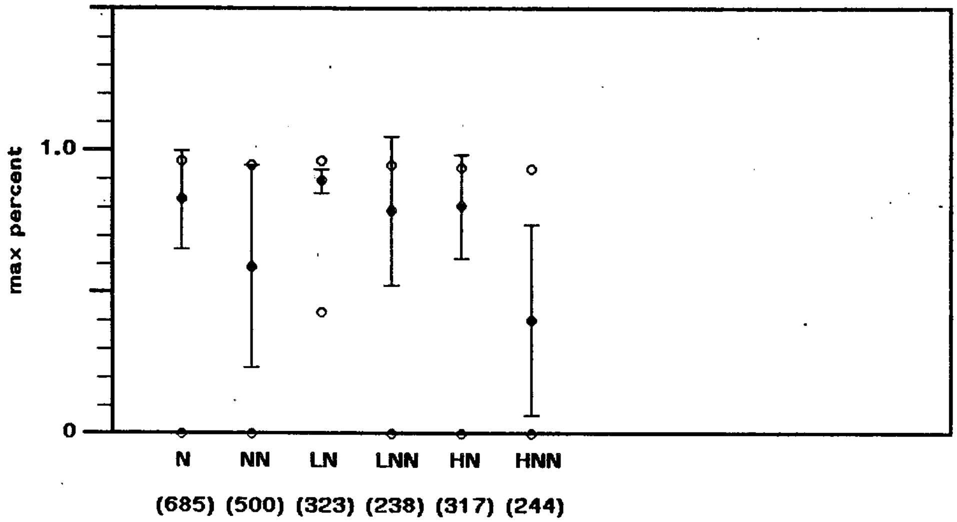


Figure 3.44: Statistics of Maximum Percent

This display summarizes the maximum percentage of nasalized and non-nasalized vowels in different contexts. From left to right they are: all nasalized vowels (N), all non-nasalized vowels (NN), nasalized low vowels (LN), non-nasalized low vowels (LNN), nasalized high vowels (HN), and non-nasalized high vowels (HNN). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

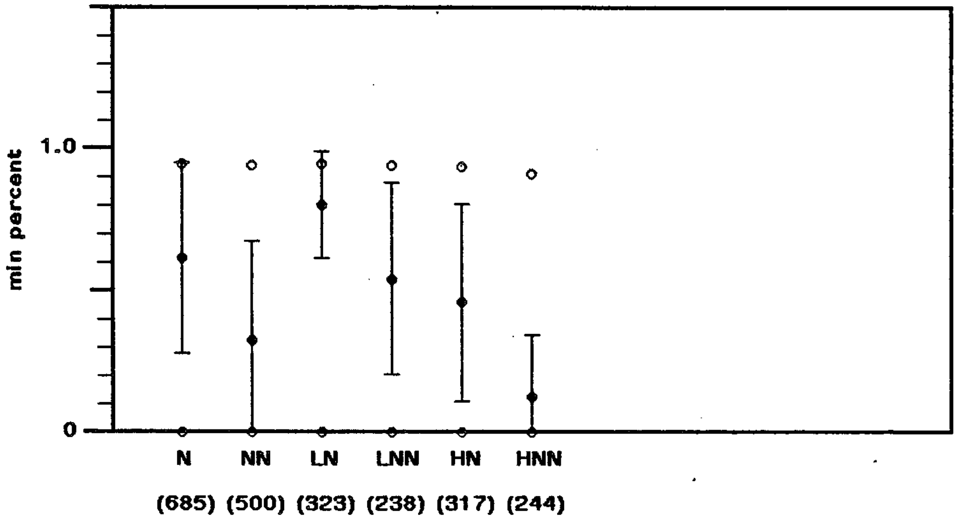


Figure 3.45: Statistics of Minimum Percent

This display summarizes the minimum percentage of nasalized and non-nasalized vowels in different contexts, which are the same as those in figure 3.44. The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open, circles display the maximum and minimum values. The numer of samples in each context are indicated below the display.

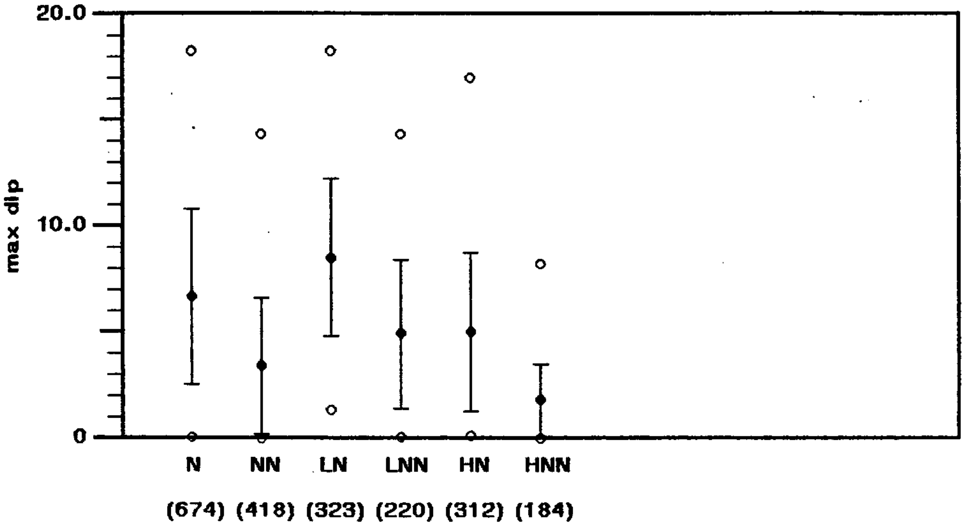


Figure 3.46: Statistics of Maximum Resonance Dip

This display summarizes the maximum resonance dip of nasalized and non-nasalized vowels in different contexts. From left to right they are: all nasalized vowels (N), all non-nasalized vowels (NN), nasalized low vowels (LN), non-nasalized low vowels (LNN), nasalized high vowels (HN), and non-nasalized high vowels (HNN). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The numer of samples in each context are indicated below the display.

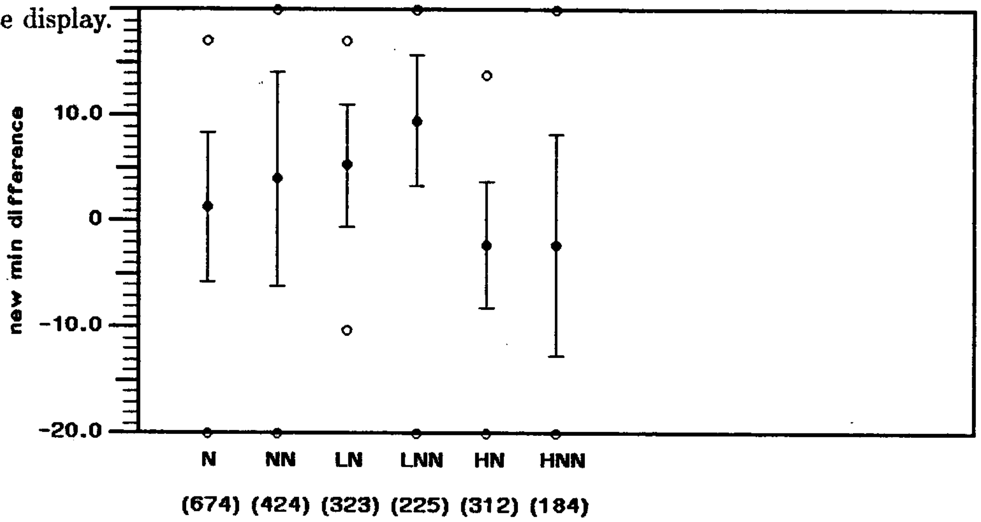


Figure 3.47: Statistics of Minimum Resonance Difference

This display summarizes the minimum resonance difference of nasalized and non-nasalized vowels in different contexts, which are the same as those in figure 3.46. The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The numer of samples in each context are indicated below the display.

Discussion

The variability of vowel spectral shapes hindered the study of nasalization. Since the main area of interest was in the first resonance region, the difficulties lay with making sure that the second formant never influenced the computations. This was naturally difficult if analysis ranges were to be kept high enough to include all of the first formants of low vowels, and kept low enough to exclude all second formants of back vowels. Since this boundary is not a fixed threshold, there are bound to be some cases where the measurement algorithms do not work correctly, as has been pointed out.

In spite of these tokens, whose main contribution was to add noise to the distributions, the acoustic study established several useful measures of nasality. The most robust measure of nasalization is the addition of an extra resonance in the low frequency region. As a result, energy in the first resonance region is more spread out, as indicated by the measure of standard deviation.

Also apparent from the acoustic study is that it is possible to discern relative degrees of nasalization by measuring the strength of the extra resonance frequency relative to the first resonance, and by measuring the amount of time that it is present in the vowel.

These observations are consistent with those made by other researchers in the past [20], [21], who have noted the presence of an extra resonance in nasalized vowels. The low resonance has been typically measured around 250 Hz. For high vowels such as /i/, the extra resonance has been observed to be located above the first formant, as was the case in this analysis. Hawkins and Stevens have suggested that nasal coupling introduces a pole zero pair into the first resonance region, and suggest that it is the zero which is the main cause of amplitude reduction of the first formant. The presence of a zero in between the low resonance and the first formant was also noted by Hattori et al., and might explain the effectiveness of the spectral dip measure of this analysis.

There are other, secondary characteristics of nasality which have been reported previously and which have not been quantified in this analysis. Most of these properties, such as observations of higher formant motion, or additional nasal resonances, were either not observed in this data, or where too difficult to attempt to extract automatically. For instance, quite often, high front vowels will exhibit another extra resonance in between the first two formants. Extracting this property however, would probably require some form of formant tracking, a difficult task in itself. Thus, this characteristic was not quantified.

Finally, it is worth examining the fact that female vowels exhibit a low resonance irrespective of context. Since these vowels are not all nasalized, there must be some other explanation for their presence. A previous study of vowels has shown that the first harmonic is enhanced when vowels have a breathy quality [2]. Since the pitch of female speakers is quite often found in the 200 to 300 Hz range, the low resonance could easily be a measure of breathiness in female speech. Although the presence of this resonance reduces the usefulness of the percentage measure for female speakers, identification of nasalized vowels is still possible, since the low resonance is strengthened when a vowel is nasalized.

3.3 Chapter Summary

The following points were established in this chapter:

1. The most robust acoustic property of a nasal consonant is a steady, low frequency resonance, which dominates the spectrum. The resonance is characterized by temporal and spectral stabilty, and by its local relative strength, properties which were quantified by the measure of low resonance percentage, and low resonance height, respectively.

2. The most robust acoustic property of a nasalized vowel is the presence of an extra resonance in the first formant region. Depending on the height of the

vowel, the extra resonance may appear above, or below the first formant. Even if the extra resonance may not be resolved from the first formant, the first resonance region is more spread out when a vowel is nasalized, a property which was quantified by the measure of standard deviation.

3. It is possiblé to discern relative degrees of nasalization by measuring the strength of the extra resonance relative to the first formant, and by measuring the amount of time that it is present in the vowel.

Chapter 4

Recognition Experiments

After observing the acoustic characteristics of nasal consonants and nasalized vowels, preliminary investigations were initiated to evaluate the potential use of these properties in speech recognition. A detailed description of the experiments that were conducted on nasal consonant and nasalized vowel detection are presented in this chapter. These experiments cannot realistically simulate a true test environment since the evaluations were made on the same database as the acoustic study. However, they do provide an indication of their potential for use in speaker-independent, speech recognition systems.

4.1 The Task

There are many different ways of restricting the problem of speaker-independent, continuous-speech, automatic nasal consonant recognition. Since the acoustic measures developed in the acoustic study were designed for the discrimination, the recognition task was structured as an identification problem. In a typical scenario, the nasal consonant detection system is given a test token and training data. The system must then classify the token as either a nasal consonant or an impostor sound. The nasalized vowel detection system must classify a test token as either next to a nasal consonant (nasalized), or not next to a nasal consonant (non-nasalized). Note that the evaluation procedure of the nasalized vowel detection system is not a true judge of nasalization, since some vowels in a non-nasal context will be nasalized, while some vowels in a nasal context will hardly be nasalized at all. A better evaluation measure would be to compare system decisions with those of human listeners.

Structuring the task in this format simplifies the problem, since it eliminates the need to detect the boundaries of the nasal consonant or vowel. These systems might be considered as specialized modules of a recognition system which are called upon only in situations which require their expertise.

4.2 The Strategy

The acoustic study quantified several parameters which characterize nasal consonants and nasalized vowels, and may discriminate them from similar sounds. Thus, it is reasonable to incorporate these measurements into detection systems for the task in hand. A given test token is then associated with a set of n values, corresponding to a set of acoustic measurements made on the test token. If we consider the set of values as a vector in an n-dimensional space, we are faced with a multidimensional decision making problem.

Multivariate decision making becomes a straightforward process when each of the parameters involved may be assumed to have jointly Gaussian distributions. In these cases, decision making is reduced to finding the distance from the test token point, to each of the normalized distributions of the possible candidates

$$D_i = (\vec{X} - \vec{m}_i)^T C_i^{-1} (\vec{X} - \vec{m}_i) \tag{4.1}$$

where D_i is the distance from the test token to candidate i; $\vec{m}_i$ is the mean vector of the parameters in the ith distribution; C_i is the covariance matrix of the parameters in the ith distribution; and $\vec{X}$ is a vector of the parameter values of the test token. For nasal consonant detection, the two candidate distributions are nasal consonants or impostors sounds, and so two distances are computed. Candidate membership is dictated by the minimum distance value.

Gaussian discrimination techniques are popular since they are quite simple, and the distance metrics correspond to maximum likelihood decisions [48]. Further, many parameter distributions may be reasonably approximated by a Gaussian of some form. When this is not the case, it is quite often possible to transform the distribution (by taking logs for example) so that a Gaussian approximation becomes reasonable. The question of a joint distribution is more difficult to account for, unless the parameters can be shown to be statistically independent.

When the joint Gaussian distribution assumption is not valid, some other procedure might be superior. Another approach which is often used, is a binary tree classifier, where, at each node in the tree, a split is made according to some criterion in one of the parameters [5]. In this fashion, a tree may be constructed which separates the data into categories without making assumptions about the underlying distributions of the parameters. Decisions are made at the bottom level of the tree based on a majority rule of the training data. Thus, if a test token happens to end up in a slot where 20 tokens of type A and 5 tokens of type B were observed during training, the test token would be classified as type A with a score of 0.8.

Although the tree classifier is attractive in the sense that it makes no assumptions of underlying distributions, it suffers from the fact that decision thresholds and actual tree structures can vary substantially, depending on the training data used.

An alternative way to combine the data would be to evaluate each parameter individually, and combine the scores at the last stage to establish some overall decision. For a binary decision (nasal or impostor), each parameter need only return a single value such as the log ratio of the likelihood that the token is nasal to the likelihood that the token is an impostor. This technique eliminates any potential multivariate information, and unless the parameters were statistically independent, might be expected to perform poorly. However, when the parameter distributions are inappropriate for standard Gaussian techniques, the likelihood approach was found to be more effective.

All of these approaches require some form of *a priori* knowledge of the distributions of the parameters. These are established through the use of training data provided to the systems. In the Gaussian approach, these values are used to compute means and covariances. In the binary classifier approach, they are used to establish node thresholds. In the the likelihood procedure, distributions are created so that an incoming test sample may be accorded a likelihood value. The actual distributions used were simple normalized histograms of the measurements. Bin widths of the histograms were set manually to ensure that the distributions would be reasonably shaped.

4.3 The Experiment

Systems were evaluated using the utterances of the database. For the nasal consonant detection task, data was divided into two groups, those in the nasal consonant class, and those which might be confused with nasal consonants (called the impostor class). These sounds included any phoneme which had acoustic characteristics similar to nasal consonants such as voice bars, liquids and glides, or weak voiced fricatives. For evaluation, there were 520 nasal consonant tokens, and 695 impostor tokens which included 357 semivowels, and 338 voice bars.

For the nasalized vowel task, the data was divided into similar groups, with the exception being that the test token was the vowel adjacent to either a nasal consonant or an impostor sound. For evaluation, there were 685 "nasalized" vowels and 500 "non-nasalized" vowels.

Ideally, the choice of the impostor sounds should be governed by a knowledge of perceptual errors. However, studies which have examined perceptual confusions between individual phonemes are scarce. Miller and Nicely have examined perceptual confusions in noisy,and band-limited signals [49]. While the study is of interest, since it indicates that nasal consonants are indeed confused with liquids, glides, and voiced stops, it is not possible to make a strong case for using their results, since the test environments are quite different. Therefore, the criterion for choosing impostors was governed more by acoustic similarity and recognition difficulty (from reports of other recognition systems).

Systems were evaluated using a rotational procedure. In each step, systems were allowed to train on the data from five of the six speakers in the database, and were tested on the data from the sixth speaker. This approach is the best approximation to a speaker-independent task, given the limited amount of data available. The following sections report the results of nasal consonant and nasalized vowel detection.

4.3.1 Detection of Nasal Consonants

There were five measures from the acoustic study which were incorporated into the nasal consonant detection system. These included:

1. *Total Energy.* The average amount of energy in the token.

2. *Energy Stability.* The average amount of change in energy in the middle of the token.

3. *Low Resonance Percentage.* The percentage of the time that there was a low frequency resonance below 350 Hz in the token.

4. *Low Resonance Amplitude.* The average amount of energy in the low frequency regions relative to total energy in the token.

5. *Low Resonance Height.* The average energy drop from the low frequency resonance to the regions immediately above.

Since it was unclear as to which decision strategy would yield the best results, an initial analysis was conducted to determine which of the three methods discussed previously performed the best. For simplicity, the systems were allowed to train on all of the tokens, since the goal was to measure the relative performance of the different strategies.

For the first examination of the data, a standard Gaussian technique was employed. Evaluating the data on the nasal consonants and impostors yielded a correct identification rate of 79%. The fact that this simple approach did so well was actually surprising, since many impostor distributions were non-Gaussian. Observation of these distributions indicated that many of the bi-modal distributions were effectively a sum of two rather standard distributions, one consisting mainly of voice bars, and the other of semivowels. This observation was also made by Mermelstein [48].

In an attempt to remedy this situation, the procedure was modified by separating the voice bars from the semivowels so that there were actually two impostor groups. An incoming test token would compute three distances, instead of two. If the minimum distance was to either of the impostor distributions, the token was labeled an impostor. Otherwise the test token was called a nasal. The average correct detection rate for this modified approach improved to 85%. It is of interest to note that glides and voice bars were rarely confused with each other, and that most of the errors were caused by labeling the nasal consonants impostors.

A binary tree classifier was evaluated next. Testing the binary tree classifier on the same data used for training is unfair, since it is possible to grow the tree during training, until there is but one element in each branch. Testing on the same data will naturally result in 100% accuracy. However, by restricting the tree to depths of around four nodes, detection rates of 91% were obtained. In order to test the sensitivity of the node thresholds, the tree was allowed to train on half of the data, and was tested on the other half (speakers were still mixed). In this case, the performance declined to 87%, indicating that thresholds were slightly sensitive to the data.

The final evaluation procedure summed the set of individual log likelihoods to come up with an overall nasal likelihood score. An average score of 89% was obtained. Confusions for all three approaches are summarized in table 4.1.

Table 4.1: Nasal Consonant Detection Confusions

	Gaussian		Tree		Likelihood	
	Nasal	Impostor	Nasal	Impostor	Nasal	Impostor
Nasal	70	30	86	14	94	6
Impostor	7	93	12	88	16	84

Since the log likelihood strategy performed slightly better than any other, and appeared to be a quite robust, it was evaluated again with the circular evaluation procedure described previously. For this case, the detection rate dropped to 88% (half a percentage point). The lack of significant decrease in the detection rate is encouraging, since it indicates that the acoustic parameters being extracted are reasonably speaker independent.

Discussion

Comparisons to other nasal consonant recognition systems are not valid at this stage of analysis, since the evaluation took place on the same database as the acoustic study. Once these parameters are tested on completely different database, the results will provide a better estimate of the speaker-independent capabilities of the system. Apart from this large qualification, it should be noted that there have not been many speaker-independent evaluations reported in the literature. Mermelstein probably had one of the more successful recognition systems although he trained and tested on only two male speakers [48].

From a recognition standpoint, it would be useful to establish the contribution made by each parameter to the overall decision. Indications from the binary tree classifier were, that the percentage measure was the most valuable, followed by the measure of the low resonance height. This implies that the main property of nasal consonant which distinguishes them from other sounds, is a continuous, low frequency resonance which dominates the spectra below 1000 Hz.

From a perceptual perspective, it would be interesting to know if the decisions made by this system are related at all to what humans would do given the same task. This topic is pursued further later on in this chapter.

4.3.2 Detection of Nasalized Vowels

There were six measures from the acoustic study which were incorporated into the nasalized vowel detection system. These included:

1. *Center of Mass.* The average value of the center of mass in the middle of the token.

2. *Standard Deviation.* The maximum value of the average standard deviation in the three vowel subregions.

3. *Maximum Resonance Percentage.* The maximum percentage of the time there is an extra resonance in the three vowel subregions.

4. *Minimum Resonance Percentage.* The minimum percentage of the time there is an extra resonance in the three vowel subregions.

5. *Maximum Resonance Dip.* The maximum value of the average dip between the first resonance and the extra resonance in the three vowel subregions.

6. *Minimum Resonance Difference.* The minimum value of the average difference between the first resonance and the extra resonance in the three vowel subregions.

As was the case for the nasal consonants, an initial analysis was performed to determine which of the three methods performed the best. Once again for simplicity, the systems were allowed to train on all of the tokens.

Using the standard Gaussian technique, a correct detection rate of 71% was obtained. No further progress was made with this technique, since there were no obvious ways to divide the data, as was the case for the nasal consonants.

When the binary tree classifier was trained and tested on the same data set, a nasalized vowel detection rate of 84% was achieved. However, when the tree was trained on half of the data, and tested on the other half, the correct detection rate fell to 79%, indicating that the node thresholds were quite sensitive to the data.

Using the log likelihood procedure, an average score of 78% was obtained. Confusions for all three approaches are summarized in table 4.2.

Table 4.2: Nasalized Vowel Detection Confusions

	Gaussian		Tree		Likelihood	
	Nasal	Non-nasal	Nasal	Non-nasal	Nasal	Non-nasal
Nasal	59	41	90	10	84	16
Non-nasal	9	91	36	64	30	70

Although the log likelihood procedure did not perform quite as well as the binary tree classifier in the initial analysis, it was used for the speaker independent test because it was found to be more stable than the tree classifier. Using the circular evaluation procedure, an average detection rate of 74% was obtained.

Unlike the nasal consonant detection systems, which performed uniformly across different speakers, and phonetic contexts, the performance of the nasalized vowel detection system varied substantially with the environment. In order to measure the difficulty of different contexts, a series of smaller evaluations, using the log likelihood procedure, were made on subsets of the vowels. The results of these experiments are summarized in table 4.3.

Table 4.3: Nasalized Vowel Detection

Evaluation	Detection Rate		
	Nasalized	Non Nasalized	Average
All	81	67	74
Male	83	78	81
Female	66	60	63
High	82	75	79
Low	75	63	69
Male High	82	75	79
Male Low	88	83	85
Female High	74	71	73
Female Low	56	67	61

From this data, it is clear that discrimination between nasalized and non-nasalized low vowels, spoken by female speakers, is quite difficult. It is also evident, that it is more difficult to detect nasality in the vowels of female speakers than in those of male speakers. Note that care must be taken to interpret the last four entries of the table since there were only two speakers in the training distributions.

Discussion

While 74% correct is better than chance, it leaves a large number of vowels for which no confident statement may be made about the presence of an adjacent nasal consonant. The main reason for this is that speakers nasalize to different degrees. Thus, in attempting to operate in a speaker independent environment, the individual distributions are being smeared.[1]

The deterioration of the detection scores for female speech is understandable, since there is often an extra low resonance in the sonorant regions, as illustrated for the vowel /æ/, in figure 4.1. Since, an extra resonance is a major acoustic

[1] Some earlier speaker dependent studies obtained detection rates over 10% better than those reported here.

difference between nasalized and non-nasalized vowels, it is natural to expect system performances to deteriorate when the low resonance is present in the speech signal irrespective of nasality. It is interesting to note that for female speakers, the system was able to identify nasalization in high vowels better than for low vowels. Since the low resonance of female speakers is always below the first formant, high vowels which have a nasal resonance *above* the first formant are uniquely nasal, and so, may be identified correctly.

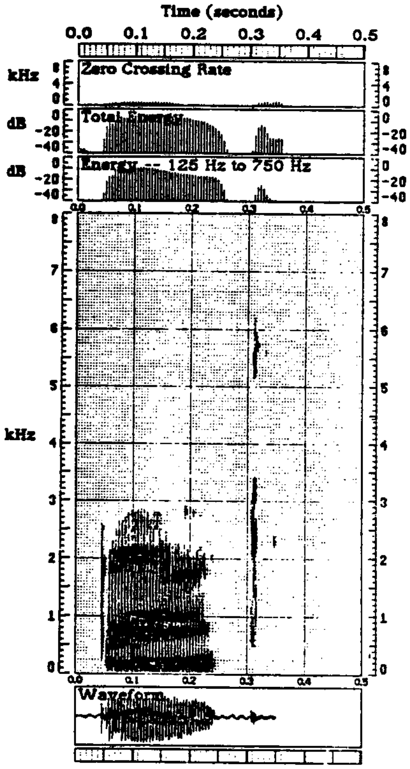


Figure 4.1: A Spectrogram of the word *back*

The performance of male speakers, is more intuitively acceptable, since the nasal resonance tends to be more "distinct" in low vowels, than in high vowels. Thus, one would expect to be able to detect nasalization more successfully in low vowels, which was confirmed in this experiment.

Of course, it is not clear how well the system is actually measuring nasalization in vowels. One way to get a better idea of this would be to perform a perceptual experiment on listeners given the same task. If the detection systems were extracting some perceptually relevant property of nasalization, then the there should some correlation between the two measures. The next section investigates this concept in more detail.

4.4 A Perceptual Evaluation

In order to provide a perceptual evaluation of the automatic detection systems , a listening experiment was performed which tried to measure people's ability to perceive nasality when all context had been stripped away. The experiment consisted of tests in which part of the speech waveform was extracted from continuous speech. For the three tests, the speech segment corresponded to:

- a murmur such as a nasal consonant, glide, or voice bar,
- a vowel adjacent to a murmur and,
- both the murmur and the adjacent vowel.

Each test consisted of forty tokens (twenty nasal and twenty non-nasal) spliced from utterances in the data base. Each token was smoothed at the ends to eliminate artifacts due to the splicing procedure, and played three times in succession. Subjects were asked to decide whether they thought the token contained a nasal (or for the second test, if the vowel was adjacent to a nasal) or a different speech sound.

The results from a panel of 20 listeners indicate that nasal consonants can be identified correctly about 65% of the time. There is some dependence on the duration of the segment but not a significant amount. Listeners were able to tell nearly 65% of the time whether a vowel was adjacent to a nasal consonant or not. Low vowels tended to be called nasal irrespectively of the presence of a nasal consonant. Listeners performed the best when they were given both the murmur and the adjacent vowel to listen to, scoring over 85%.

Comparison to Detection Systems

When the tokens of the first listening test were run on the nasal consonant detection system, 64% of the tokens were identified correctly. This was effectively the same as the listeners. The nasality scores produced by listeners and the detection system seem to be rather correlated, as shown in figure 4.2.

When the tokens of the second test were run on the nasalized vowel detection system, 74% of the tokens were correctly identified. This result is notably better than that obtained by human listeners. For this test as well, there were indications that the scores were somewhat correlated, as shown in figure 4.3.

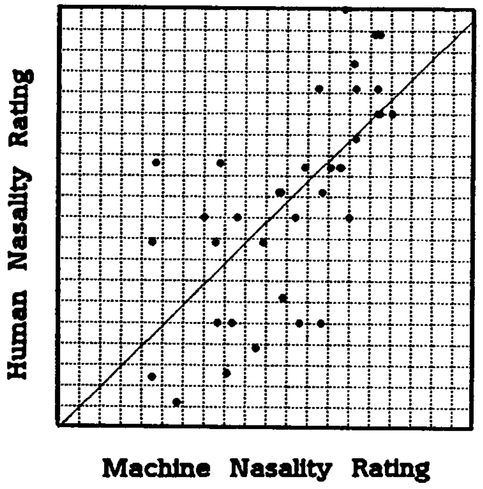


Figure 4.2: Nasality Rating for Murmur Tokens: Human versus Machine

This figure plots two measures of nasality for murmur tokens in the perceptual study. On the vertical axis are listeners nasality rating of the token. Likelihood scores given by machine are plotted on the horizontal axis.

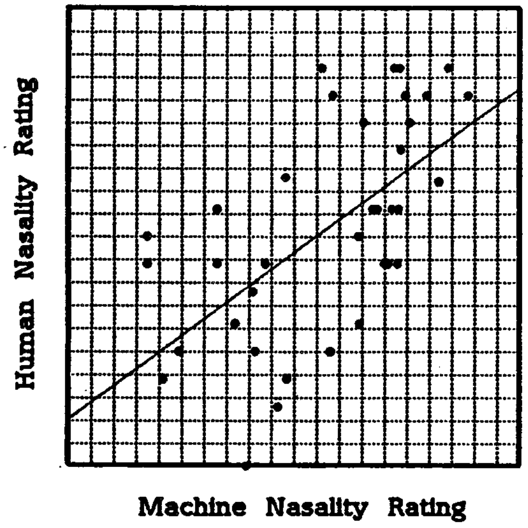


Figure 4.3: Nasality Rating for Vowel Tokens: Human versus Machine

This figure plots two measures of nasality for vowel tokens in the perceptual study. On the vertical axis are listeners nasality rating of the token. Likelihood scores given by machine are plotted on the horizontal axis.

Discussion

Since the nasal consonant detection system scored well below average on the murmur tokens used in the perceptual experiment, it is clear that these sounds were a difficult subset of the database. Thus, in general, one could expect listeners to be more successful at the nasal consonant detection task than was observed here. The fact that the results were somewhat correlated provides an indication that the nasal consonant detector is extracting relevant parameters from the speech signal.

Perhaps one of the more informative results of the vowel study was that it indicated that listeners are indeed able to use information in the vowel to detect nasal consonants. These results support the hypothesis of Ali et al that listeners use nasalization to lighten the phoneme processing load [1]. The fact that the vowel detection system performed better than listeners at this task is probably due in part to the fact that nasalized vowels have no phonetic distinction in American

English. Thus, untrained listeners were not aware of the concept of nasalization, and had a harder time detecting this property. Another reason for this difference in performance could have been that the detection system was allowed to train on utterances spoken by the same speakers, while human listeners were not.

4.5 Chapter Summary

The main points of this chapter are:

1. Using a log likelihood decision strategy employing robust measures established in the acoustic analysis, nasal consonant detection rates of 88% were obtained.

2. Using a similar decision strategy, a nasalized vowel detection rate of 74% was obtained. Detection rates varied substantially with speaker sex, and vowel height. The best decision rate of 85%, was obtained for low vowels spoken by male speakers. The worst decision rate of 61%, was obtained for low vowels spoken by female speakers.

3. A perceptual evaluation of a subset of the database indicates that system decisions tend to be correlated with decisions made by human listeners performing a similar task.

Chapter 5

Summary and Future Work

5.1 Summary

There are several conclusions which can be made from this research. First, the acoustic analysis established that nasal consonants are characterized by a low resonance, typically centered between 200 and 350 Hz, which dominates the overall spectrum. Another property of the low resonance, which was found to be typical of nasal consonants, was a sudden drop in energy at slightly higher frequencies in the first resonance region. This measure was found to be most effective in discriminating nasal consonants from semivowels. This parameter should also rule out most vowels, with the exception of some high front vowels such as /i/, or a raised /u/.

The acoustic analysis also found that the most robust measure of nasalization is the presence of an extra resonance in the low frequency region, resulting in a first resonance region where the energy is more spread out, as indicated by the measure of standard deviation. The acoustic study also established that it is possible to discern relative degrees of nasalization by measuring the relative strength of the extra resonance to the first resonance, and by measuring the amount of time that it is present in the vowel.

Finally, the preliminary investigations of nasal consonant and nasalized vowel detection provide indications that these acoustic properties are useful for applications in speaker-independent speech recognition systems.

5.2 Future Work

Although this research observed many characteristics of nasality, there are still many areas which require further investigation. One area which was only briefly examined, was the transition region between the nasal consonant and the adjacent vowel. This time interval is worthy of serious study, since it contains the most information about the place of articulation of the nasal consonants. In addition, it contains pertinent information for discriminating semivowels from consonants.

As illustrated in figure 5.1 for the word *need*, the transition region of prevocalic nasal consonants is denoted by a sudden spectral change at high frequencies, with limited formant transitions in the vowel. This information can be used to discriminate nasal consonants from semivowels. Figure 5.1 also shows an /l/ from the word *lead*, which, although having similar acoustic characteristics to a nasal consonant, may be eliminated as a potential nasal due to a lengthy second formant transition.

Another characteristic worth quantifying, is the extension of the low frequency resonance into an adjacent vowel, as has been illustrated before in figures 2.6, 3.28, and 3.29. Although this property is not apparent for all vowels, it is a very powerful indication of the presence of a nasal consonant when it exits.

From a speech recognition standpoint, there are several ways in which this work could be extended. First, it is clear that the evaluation performed in this work is inadequate since it it was based on the same database as the acoustic analysis. For a true evaluation of the parameters developed in this work, a totally different database should be used. For speaker-independence, the database should have a

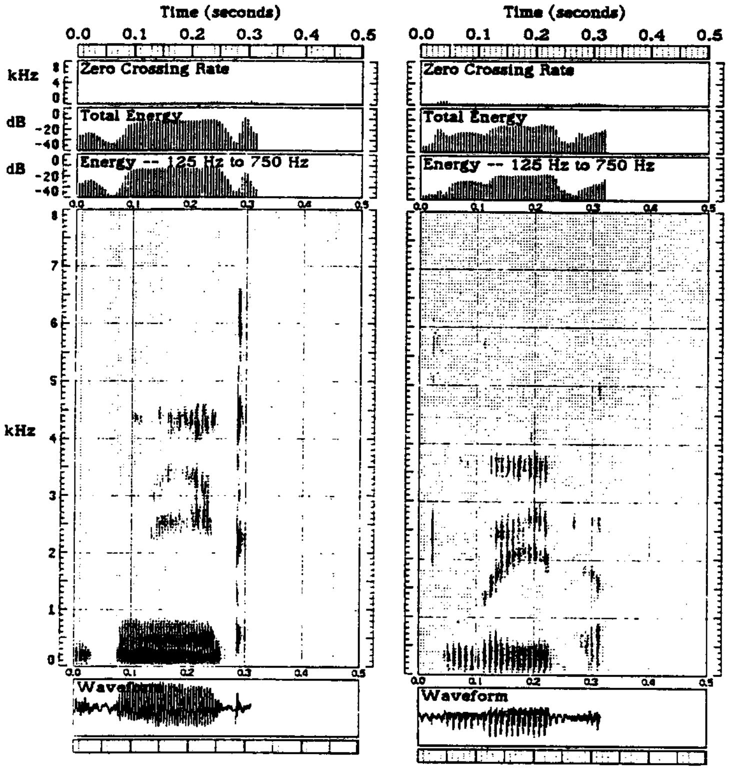


Figure 5.1: Spectrograms of the words *need*, and *lead*

large number of speakers. Since the acoustic measurements developed in the acoustic study are proposed for continuous speech, it also would be appropriate to collect a sentence database.

In order to simplify the nasal recognition problem, nasal consonant boundaries were detected manually in this research. Clearly, it would be worthwhile establishing some automatic procedure for detecting nasal consonant boundaries. A preliminary feasibility study examined the performance of a boundary detection algorithm based on locating points of maximum spectral change in the speech waveform. A similar procedure was used successfully in the past by Mermelstein [48]. An evaluation of all nasal vowel sequences in the database, the results of which are presented in figure 5.2, showed that nasal consonant boundaries can be located within 20 msec of a manually assigned boundary over 95% of the time.

The experiments in nasalized vowel detection indicated that it is possible to use

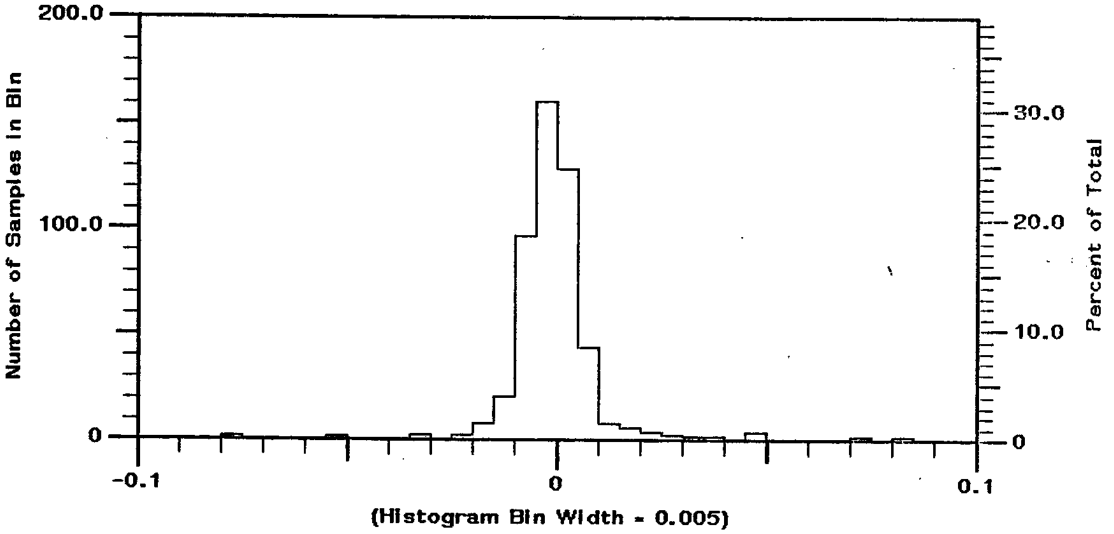


Figure 5.2: A Histogram of Boundary Detection Error

This figure contains a histogram of the errors of the automatic nasal consonant boundary detection algorithm. The error is calculated by taking the time difference between the hypothesized boundary, and a manually assigned boundary. Values are in seconds.

information in the vowel to predict the presence of a nasal consonant. Thus, it would be worthwhile to establish if this information actually improves nasal detection systems. Finally, it would be interesting to examine the usefulness of speaker adaptation in the nasalized vowel detection system, since vowel nasalization was found to be strongly speaker dependent.

Bibliography

[1] Ali, A., Gallagher, T., Goldstein, J., Daniloff, R., "Perception of Coarticulated Nasality, *Journal of the Acoustical Society of America*, Vol. 49, No. 2, pp. 538-540, 1971.

[2] Bickley, C.A., "Acoustic Analysis and Perception of Breathy Vowels", *Working Papers, Speech Communication Group*, Research Laboratory of Electronics, MIT, Vol. 1, pp. 71-82, 1982.

[3] Blomberg, M., Carlson, R., Elenius, K., Granstrom, B., "Auditory Models in Isolated Word Recognition, *Proceedings ICASSP 84*, San Diego, CA, pp. 17.9.1-17.9.4, 1984.

[4] Blumstein, S.E., Stevens, K.N., "Acoustic Invariance in Speech Production: Evidence from Measurements of the Spectral Characteristics of Stop Consonants, *Journal of the Acoustical Society of America*, Vol. 66, No. 4, pp. 1001-1017, 1979.

[5] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 1984.

[6] Bush, M.A., Kopec, G.E., Zue, V.W., "Selecting Acoustic Features for Stop Consonant Identification, *Proceedings of ICASSP 83*, Boston, MA, 1983, pp. 742-745.

[7] Cole, R.A., Editor *Perception and Production of Fluent Speech*, Lawrence-Erlbaum Ass., Hillsdale, New Jersey, 1980.

[8] Dautrich, B.A., Rabiner, L.R., Martin, T.B., "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition, *IEEE Transactions ASSP*, Vol. 31, No. 4, pp. 793-806, 1983.

[9] De Mori, R., Gubrynowicz, R., Laface, P., "Inference of a Knowledge Source for the Recognition of Nasals in Continuous Speech, *IEEE Transactions ASSP*, Vol. 5, pp. 538-549, 1979.

[10] Dickson, D.R., "Acoustic Study of Nasality, *Journal of Speech and Hearing Research*, Vol. 5, No. 2, pp. 103-111, 1962.

[11] Dixson, N.R., Silverman, H.F., "A General Language Operated Decision Implementation System (GLODIS): Its Application to Continuous-Speech Segmentation, *IEEE Transactions ASSP*, Vol. 24, pp. 137-162, 1976.

[12] Fairbanks, G., House, A.S., Stevens, A.L., "An Experimental Study of Vowel Intensities, *Journal of the Acoustical Society of America,* Vol. 22, No. 4, pp. 457-459.

[13] Fant, G., *Acoustic Theory of Speech Production*, Mouton and Co., 's-Gravenhage, Netherlands, 1960.

[14] Flanagan, J.L., *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, 1972.

[15] Fujimura, O., "Spectra of Nasalized Vowels, *Research Laboratory of Electronics, MIT Quarterly Report*, No. 58., pp. 214-218, 1960.

[16] Fujimura, O., "Analysis of Nasal Consonants, *Journal of the Acoustical Society of America*, Vol. 34, No. 12, pp. 1865-1875, 1962.

[17] Fujimura, O., "Formant-Antiformant Stucture of Nasal Murmers, *Proceedings of the Speech Communication Seminar*, Vol 1, Stockholm: Royal Institute of Technology, Speech Transmission Laboratory, pp. 1-9, 1962.

[18] Fujimura, O., Lindqvist, J., "Sweep-Tone Measurements of the Vocal Tract Characteristics, *Journal of the Acoustical Society of America*, Vol. 49, No. 2, pp. 541-558, 1971.

[19] Gillmann, R.A., "Automatic Recognition of Nasal Phonemes, *Proceedings IEEE Symposium on Speech Recognition*, Pittsburgh, PA, 1974, pp. 74-79.

[20] Hattori, S., Yamamoto, K., Fujimura, O., "Nasalization of Vowels in Relation to Nasals, *Journal of the Acoustical Society of America*, Vol. 30, No. 4, pp. 267-274, 1958.

[21] Hawkins, S., Stevens, K.N., "A cross-language study of the perception of nasal vowels, Paper presented at the 105th meeting of the Acoustical Society of America, Cincinatti, Ohio, 1983.

[22] Hess, W.J., "A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speech, *IEEE Transactions ASSP*, Vol. 24, pp 14-25, 1976.

[23] House, A.S., Stevens, K.N., "Analog Studies of the Nasalization of Vowels, *Journal of Speech and Hearing Disorders*, Vol. 22, No. 2, pp. 218-232, 1956.

[24] House, A.S., "Analog Studies of Nasal Consonants, *Journal of Speech and Hearing Disorders*, Vol. 22, pp. 190-204, 1957.

[25] Hyde, S.R., "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature, *Human Communication: A Unified View*, edited by E.E. David and P.B. Denes, McGraw-Hill, New York, 1972.

[26] Jakobson, R., Fant, G., Halle, M., *Preliminaries to Speech Analysis*, MIT Press, Cambridge, Mass, 1963.

[27] Kawasaki, H., "The perceived nasality of vowels with gradual attenuation of adjacent nasal consonants, Paper presented at joint meeting of the Acoustical Society of America and Acoustical Society of Japan, Honolulu, HI, 1978.

[28] Klatt, D.H., "Review of the ARPA Speech Understanding Project, *Journal of the Acoustical Society of America*, Vol. 62, No. 6, pp. 1345-1366, 1977.

[29] Klatt, D.H., "Speech Perception: a Model of Acoustic-Phonetic Processing and Lexical Access, *Journal of Phonetics*, Vol. 7, pp. 279-312, 1979.

[30] Kopec, G.E., "Voiceless Stop Consonant Identification Using LPC Spectra, *Proceedings of ICASSP 84*, San Diego, CA, 1984.

[31] Kurowski, K., Blumstein, S.E., "Perceptual Integration of the Murmur and Formant Transitions for Place of Articulation in Nasal Consonants, *Journal of the Acoustical Society of America*, Vol. 76, No. 2, pp. 383-390, 1984.

[32] Labov, W., Yaeger, M., Steiner, R., "A Quantative Study of Sound Change in Progress", Report on National Science Foundation Contract, NSF-65-3287, University of Pennsylvania, 1972.

[33] Larkey, L.S., Wald, J., Strange, W., "Perception of synthetic nasal consonants in initial and final syllable position, *Perception and Psychophysics*, Vol. 23, No. 4, pp. 299-312, 1978.

[34] Lea, W.A., Editor *Trends is Speech Recognition*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1980.

[35] Leung, H.C., Zue, V.W., "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech, *Proceedings of ICASSP 84*, San Diego, CA, 1984.

[36] Liberman, A.M., Delattre, P., Cooper, F.S., Gerstman, L.J., "The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants, *Psychological Monographs*, Vol. 68, No. 8, pp. 1-13, 1954.

[37] Lindqvist, J., Sundberg, J., (1972), "Acoustic Properties of the Nasal Tract, *Speech Transmission Laboratory Quarterly Progress Status Report*, No. 1, Royal Institute of Technology, Stockholm, pp. 13-17, 1972.

[38] Lintz, L.B., Sherman, D., "Phonetic Elements and Perception of Nasality, *Journal of Speech and Hearing Research*, Vol. 4, No. 4, pp. 381-396, 1961.

[39] Maeda, S., "The Role of the Sinus Cavities in the Production of Nasal Vowels, *Proceedings of ICASSP 82*, Paris, France, 1982, pp. 911-914.

[40] Maeda, S., "Acoustic Correlates of Vowel Nasalization: a Simulation Study, Paper presented at the 104th meeting of Acoust. Soc. Amer., Orlando, FL, 1982.

[41] Makhoul, J.I., Wolf, J.J., "Linear Prediction and the Spectral Analysis of Speech, Bolt, Beranek and Newman Report No. 2304, 1972.

[42] Makhoul, J.I., "Linear Prediction: A Tutorial Review, *Proc. IEEE*, Vol. 63, pp. 561-580, April 1975.

[43] Malécot, A., "Acoustic Cues for Nasal Consonants: An Experimental Study Involving a Tape-Splicing Technique, *Language*, Vol. 32, pp. 274-284, 1956.

[44] Malécot, A., "Vowel Nasality as a Distinctive Feature in American English, *Language*, Vol. 36, No. 2, pp. 222-229, 1960.

[45] Markel, J.D., Gray Jr., A.H., *Linear Prediction of Speech*, Springer-Verlag, New-York, 1976.

[46] Mártony, J., "The role of formant amplitudes in synthesis of nasal consonants, *Speech Transmission Laboratory Quarterly Progress Status Report*, No. 3, Royal Institute of Technology, Stockholm, pp. 28-31, 1964.

[47] Mathews, M.V., Miller, J.E., David Jr., E.E., "Pitch Synchronous Analysis of Voiced Sounds, *Journal of the Acoustical Society of America*, Vol. 33, No. 2, pp. 179-186, 1961.

[48] Mermelstein, P., "On Detecting Nasals in Continuous Speech, *Journal of the Acoustical Society of America*, Vol. 61, No. 2, pp. 581-587, 1977.

[49] Miller, G.A., Nicely, P.E., "An Analysis of Perceptual Confusions Among Some English Consonants, *Journal of the Acoustical Society of America*, Vol. 27, No. 2, pp. 338-352, 1955.

[50] Nakata, K., "Synthesis and Perception of Nasal Consonants, *Journal of the Acoustical Society America*, Vol. 31, No. 6, pp. 661-666, 1959.

[51] Nguyen, D.T., Guern, B., "Effects of Nasal Coupling on the Vowels, Paper presented at the 99th meeting of the Acoustical Society of America, Atlanta, GA, 1980.

[52] Noll, A.M., "Cepstrum Pitch Determination, *Journal of the Acoustical Society of America*, Vol. 41, No. 2, pp. 293-309, 1967.

[53] Nord, L., "Experiments with Nasal Synthesis, *Speech Transmission Laboratory Quarterly Progress Status Report*, No. 2-3, Royal Institute of Technology, Stockholm, pp. 14-19, 1976.

[54] Nord, L., "Perceptual Experiments with Nasals, *Speech Transmission Laboratory Quarterly Progress Status Report*, No. 2-3, Royal Institute of Technology, Stockholm, pp. 5-8, 1976.

[55] Oppenheim, A.V., "Speech analysis-synthesis system based on homomorphic filtering, *Journal of the Acoustical Society of America*, Vol. 45, No. 2, pp. 458-465, 1969.

[56] Oppenheim A.V., Schafer, R.W., *Digital Signal Processing*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1975.

[57] Oppenheim A.V., *Applications of Digital Signal Processing*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[58] Pinson, E.N., "Pitch Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths, *Journal of the Acoustical Society of America*, Vol. 35, No. 8, pp. 1264-1273, 1963.

[59] Peterson, G.E., Barney, H.L., "Control Methods Used in a Study of the Vowels, *Journal of the Acoustical Society of America*, Vol. 24, No. 2, pp. 175-185, 1952.

[60] Pols, L.C., Tromp, H.R.C., Plomp, R., "Frequency Analysis of Dutch Vowels from 50 Male Speakers, *Journal of the Acoustical Society of America*, Vol. 53, No. 4, pp. 1093-1101, 1973.

[61] Rabiner, L.R., Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[62] Raphael, L., Dormann, M., Freeman, F., "Vowel and Nasal Duration as Cues to Voicing in Word-Final Stop Consonants: Spectrographic and Perceptual Studies, *Journal of Speech and Hearing Research*, Vol. 18, pp. 839-400, 1975.

[63] Recasens, D., "Place Cues for Nasal Consonants with special reference to Catalan, *Journal of the Acoustical Society of America*, Vol. 73, No. 4, pp. 1346-1353, 1983.

[64] Repp, B.H., "Perception of the [m]-[n] Distinction: Insights from Four Converging Procedures, Paper presented at 108th meeting of the Acoustical Society of America, Minneapolis, Minnesota, 1984.

[65] Rosenberg, A.E., "Effect of Glottal Pulse Shape on the Quality of Natural Vowels, *Journal of the Acoustical Society of America*, Vol. 49, No. 2, pp. 583-590, 1971.

[66] Schafer, R.W., Rabiner, L.R., "System for automatic formant analysis of voiced speech, *Journal of the Acoustical Society of America*, Vol. 47, No. 2, pp. 634-648, 1970.

[67] Searle, C.L., Jacobson, J.Z., Rayment, S.G., "Stop consonant discrimination based on human audition, *Journal of the Acoustical Society of America*, Vol. 65, No. 3, pp. 799-809, 1979.

[68] Seneff, S., Klatt, D.H., Zue, V.W., "Design considerations for optimizing the intelligiblity of DFT-based, pitched-excited, critical-band spectrum speech analysis/resynthesis system, *Speech Communication Group Working Papers*, No. 1, pp. 31-46, 1982.

[69] Shipman, D.W., "SpireX: Statistical Analysis in the Spire Acoustic-Phonetic Workstation, Proceedings of ICASSP 83, Boston, MA, pp. 1360-1363, 1983.

[70] Stevens, K.N., Klatt, M., "Study of Acoustic Properties of Speech Sounds, Bolt, Beranek and Newman Report No. 1669, 1968.

[71] Stevens, K.N., "Study of Acoustic Properties of Speech Sounds II, and Some Remarks on the Use of Acoustic Data in Schemes for Machine Recognition of Speech, Bolt, Beranek and Newman Report No. 1871, 1969.

[72] Su, L.S., Li, K.P., Fu, K.S., "Identification of speakers by the use of nasal coarticulation, *Journal of the Acoustical Society of America*, Vol. 56, No. 6, pp. 1876-1882, 1974.

[73] Weinstein, C.J., McCandless, S.S., Mondshein, L.F., Zue, V.W., "A System for Acoustic-Phonetic Analysis of Continuous Speech, *IEEE Transactions ASSP*, Vol. 23, pp. 54-67, 1975.

[74] White, G.M., Neely, R.B., "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming, *IEEE Transactions ASSP*, Vol. 24, No. 2, pp. 183-188, 1976.

[75] Wright, J., "The Behavior of Nasalized Vowels in the Perceptual Vowel Space, *Report of the Phonology Laboratory*, No. 5, Berkely, CA, 1980.

[76] Zue, V.W., "Acoustic Characteristics of Stop Consonants: A Controlled Study, Sc.D. Thesis, M.I.T., 1976.

[77] Zue, V.W., Laferriere, M., "Acoustic Study of Medial /t,d/ in American English, *Journal of the Acoustical Society of America*, Vol. 66, No. 4, pp. 1039-1050, 1979.

[78] Zue, V.W., "Acousic-Phonetic Knowledge Representation: Implications from Spectrogram Reading Experiments, *Proceedings of the 1981 NATO Advanced Summer Institute on Automatic Speech Analysis and Recognition*, P. Reidel Publishing Co., 1981.

[79] Zue, V.W., Sia, E.B., "Nasal Articulation in Homorganic Clusters in American English, Paper presented at 102nd meeting of the Acoustical Society of America, Miami, FL, 1981.

[80] Zwicker, E., Terhardt, E., Paulus, E., "Automatic speech recognition using psychoacoustic models, *Journal of the Acoustical Society of America*, Vol. 65, No. 2, pp. 487-498, 1979.

Appendix A

Corpus Words

The corpus has a total of 203 different words containing nasal consonants in various positions, both as a single consonant and as part of a cluster. Care was taken to include words that formed minimal pairs, as well as words with acoustic characteristics that are similar to nasals. The following tables section the corpus words into their basic phonetic environment.

Table A.1: Consonant Nasal Clusters

m	n	w/o nasal	across syllable		similar words				
smack	snack	sack			slack	nack	mack	lack	back
smoke	snowed	sewed	gismo		slowed	note	low	moat	dote
smock	snot	sought			mod	slot	lot	not	swat
smitten	snit	sit	ethnic	parsnip	slit	lit	knit	mitt	
film	kiln	kill							
dorm	corn	whorl							

Table A.2: Nasal Stop Consonant Clusters

with stop		w/o nasal		w/o stop	across syllable		similar words		
camp		cap	cab	cam	camper	campbell	can		
sump		sup	sub	some	somebody		sun	sung	
font	fond	fought		fawn	fondest		fall	fault	
bent	bend	bet	bed	ben	bending	sentry	bell	belt	
pant	panned	pat	pad	pan	panter	pander	pal	pam	
sink		sick		sing	sinking	single	sill	sin	silk
sunk		suck		sung	sunken	hunger	sun	some	sulk

Table A.3: Syllable Initial Nasals

m	n	similar words			across syllable		
made	nape	bade	tape		helpmate	cognate	
mitt	nip	bit	dip	lip	abnegate	admit	picnic
meat	need	beat	deed	lead	voltmeter	technique	
mack	nack	back	tack	lack	enigma		
moat	note	boat	dote	vote	utmost	ignore	
mutt	nut	but	dud		chipmunk	pignut	

Table A.4: Syllable Final Nasals

m	n	ŋ	similar words			across syllable	
cam	can		cab	cad		campbell	
dim	din	ding	dip	did	dill	skimpy	
some	sun	sung	sub	sud		sunken	somebody
comb	bone		cope	bode	bowl	lonely	homely

Table A.5: Nasal Fricative Clusters

w/o fricative	with fricative		across syllable	similar words	
warm	warms	warmth		worn	
	triumph		triumphant		
limb	lymph				
won	once	ones			
pin	pinch	pins	pinching	pill	pills
strain	strange	strains	stranger		
string	strength	strings			

Table A.6: Intervocalic Nasals

m	n	ŋ	similar words	
simmer	sinner	singer	tiller	critter
hammer	banner	hanger	matter	
rummy	runny		ruddy	sully
comic	conic		polish	
demise	denies		relies	devise
hammock	bannock		haddock	havock

Table A.7: Syllabic Nasals

m	n	similar word
bottom	button	bottle
totem	oaten	total

Table A.8: Miscellaneous

chimney
inmate
hangman
omnibus
damnation
dalmation
arsenic
decimal
animal
flannel

Appendix B

Phonetic Transcription Alignment Procedures

In order to be able to analyze utterances with the SpireX statistical package, time aligned phonetic transcriptions were required. This appendix describes the procedure for time alignment, and summarizes the rules used.

In this research, the phonetic transcriptions were aligned manually to the waveform using the Spire facility available on MIT Lisp Machine work stations. A typical transcription layout, illustrated in figure B.1, contains:

1. the orthographic, and phonetic transcriptions,
2. a menu of possible phonetic symbols,
3. a broad-band spectrogram of the utterance,
4. one compressed, and one expanded view of the speech waveform and,
5. a short-time spectral slice, computed with a 6.6 msec hamming window, at the position of a time cursor.

By positioning a cursor and a marker, a segment region may be established. As shown in figure B.1, this segment region is indicated by two vertical lines on the spectrogram, or speech waveforms. A phonetic symbol is associated with this time segment by selecting an element from the symbol table.

The time alignment process usually proceeds from left to right. In a typical alignment operation, the spectrogram is used to position the time cursor near the next segment boundary. The exact position of the boundary is determined by observation of the expanded speech waveform, the short-time spectra, and when necessary, by listening to the speech segment.

The boundary between a nasal consonant and an adjacent sonorant is not difficult to establish since it is denoted by sudden spectral and intensity changes. This is reflected by sharp changes in the spectrogram, as shown in figure B.2 for the word *simmer*. Note that the periodic waveform changes its shape noticeably on either side of the boundary. For these types of transitions the boundary was set at the point of maximum spectral change.

The boundary between a nasal consonant and a voiceless obstruent, or period of silence, was set at the onset (or offset) of voicing in the waveform. Figure B.3 illustrates the case for the fricative nasal cluster in the word *smack*. Here, the boundary was set at the onset of voicing of the nasal consonant. The period of epenthetic silence in between the fricative and the nasal consonant is caused by an asynchrony mistiming of the movements of the articulators, and is common in all fricative nasal clusters.

The boundary between a nasal consonant and a voiced obstruent was determined by an onset of some other acoustic characteristic. Figure B.4 illustrates the case of the word *warms*, where the boundary is set at the onset of frication. Figure B.5 illustrates the case of the word *bending*, where the boundary was determined by the lack of energy immediately above the low frequency resonance (relative to the rest of the murmur, which was labeled as the nasal consonant). Note that this difference is quite subtle.

The most difficult boundaries to establish were between nasal consonants and voice bars, as shown in the previous figure, or between two different nasal consonants, as illustrated in figure B.6, for the word *inmate*. In these cases, the boundary was determined by observing subtle changes in the short-time spectra, and by listening to the individual regions in the speech waveform.

In many intervocalic environments, the consonant /n/ was produced as a nasal flap, as shown in figure B.7, for the word *bannock*. For transcription purposes, any /n/, in an intervocalic environment, which was less than two pitch periods long, was labeled a nasal flap.

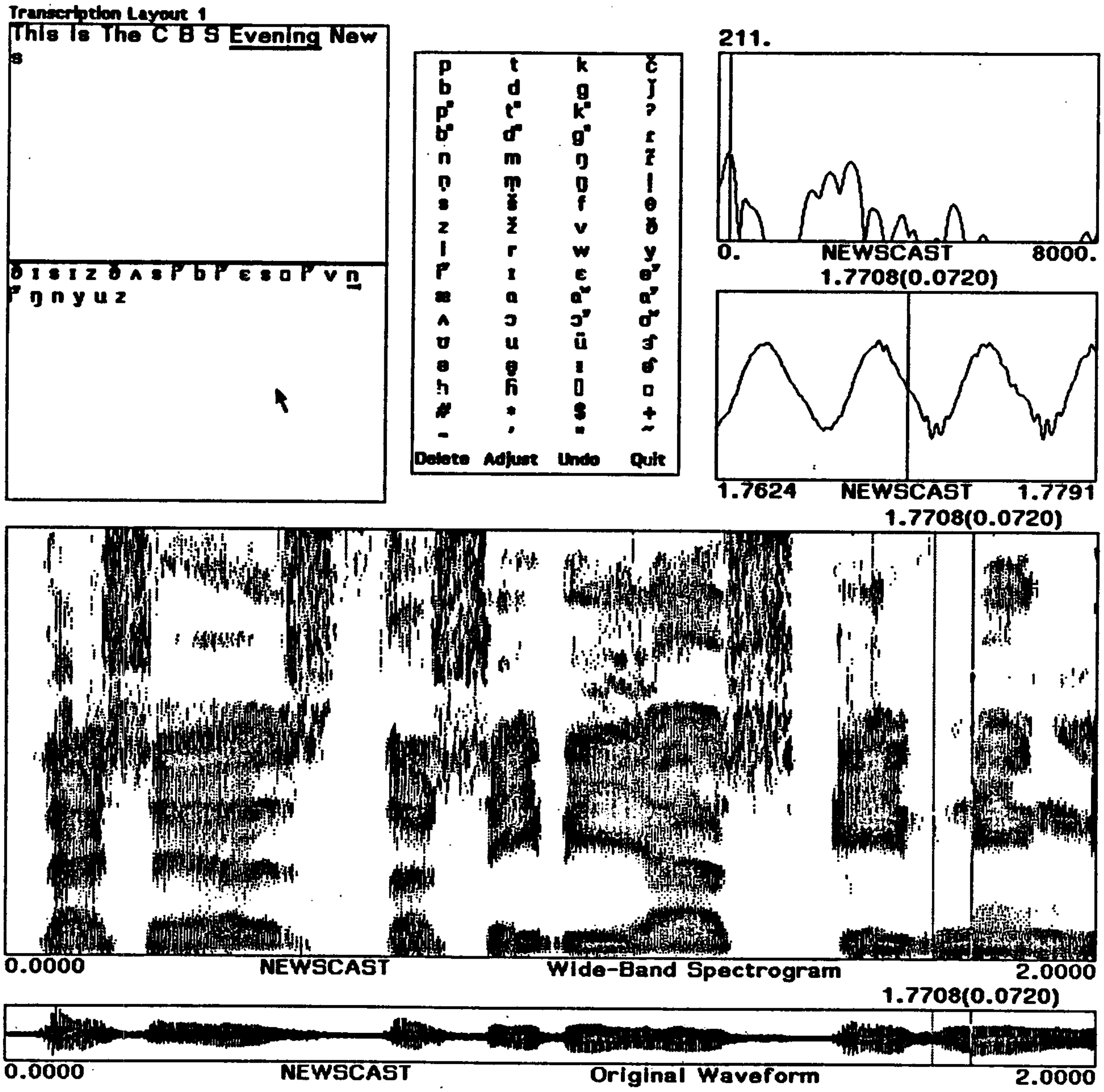
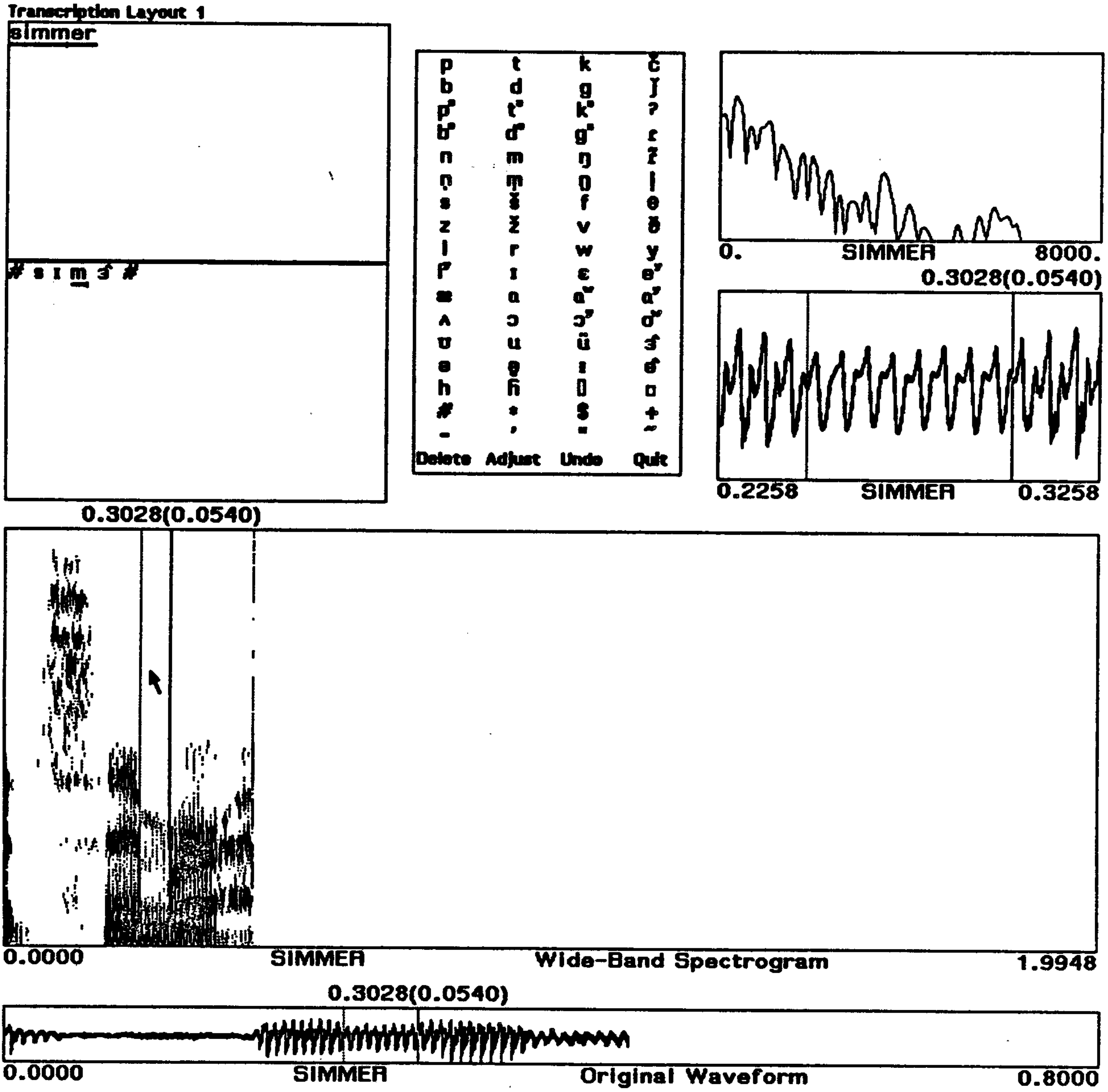


Figure B.1: The Spire Phonetic Transcription Layout

This figure contains a typical transcription layout of the Spire facility. The layout contains, counter-clockwise from the upper left, the orthographic transcription, the phonetic transcription, a broad-band spectrogram, the speech waveform, an expanded speech waveform, and a short-time spectra computed at the time of the cursor. A phonetic symbol table is located in the middle of the layout. The cursor is indicated by a vertical line in the spectrogram, or speech waveform, and is controlled by the mouse.



simmer

s ɪ m ɝ

Figure B.2: The Transcription of the word *simmer*

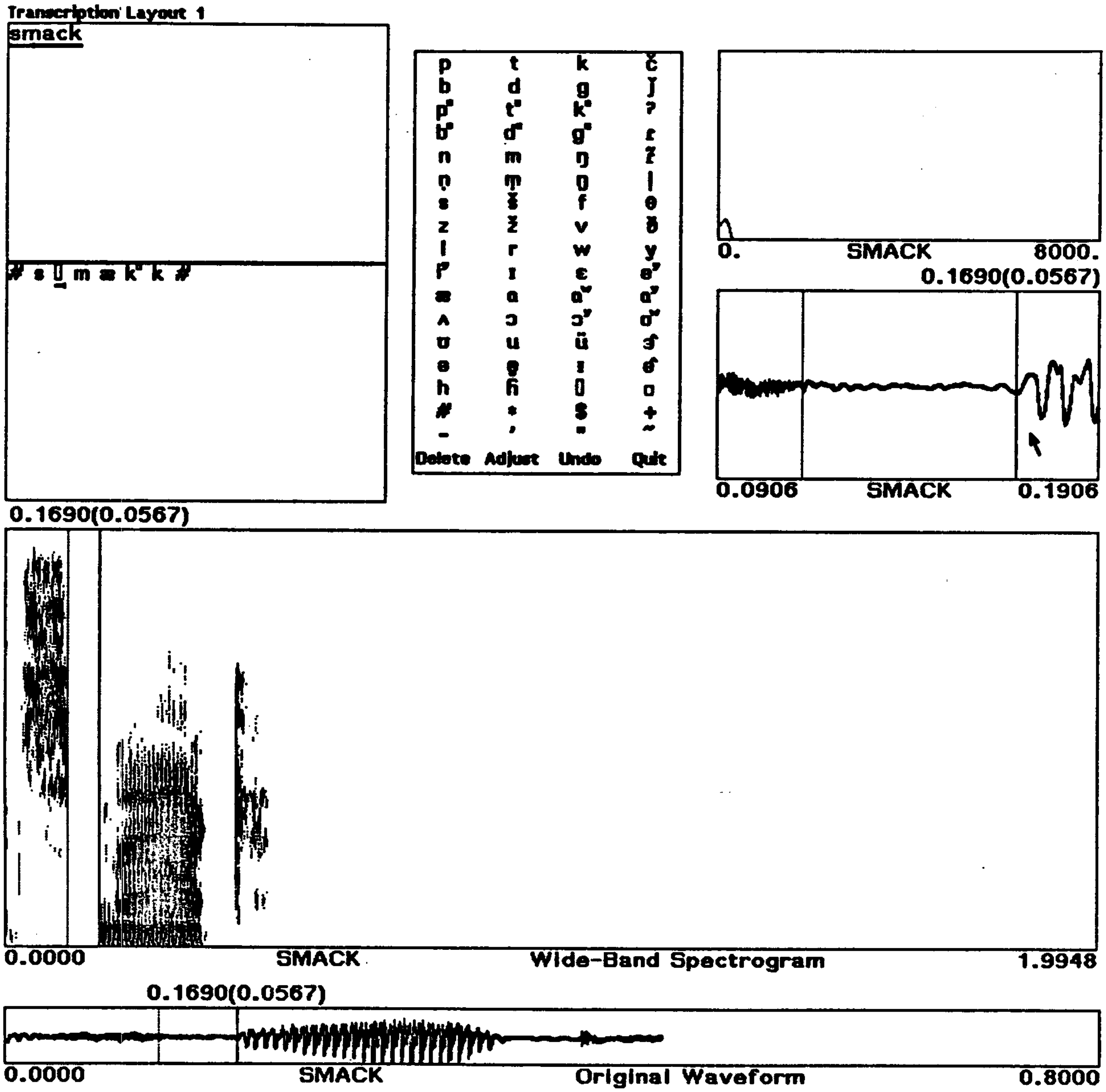


Figure B.3: The Transcription of the word *smack*

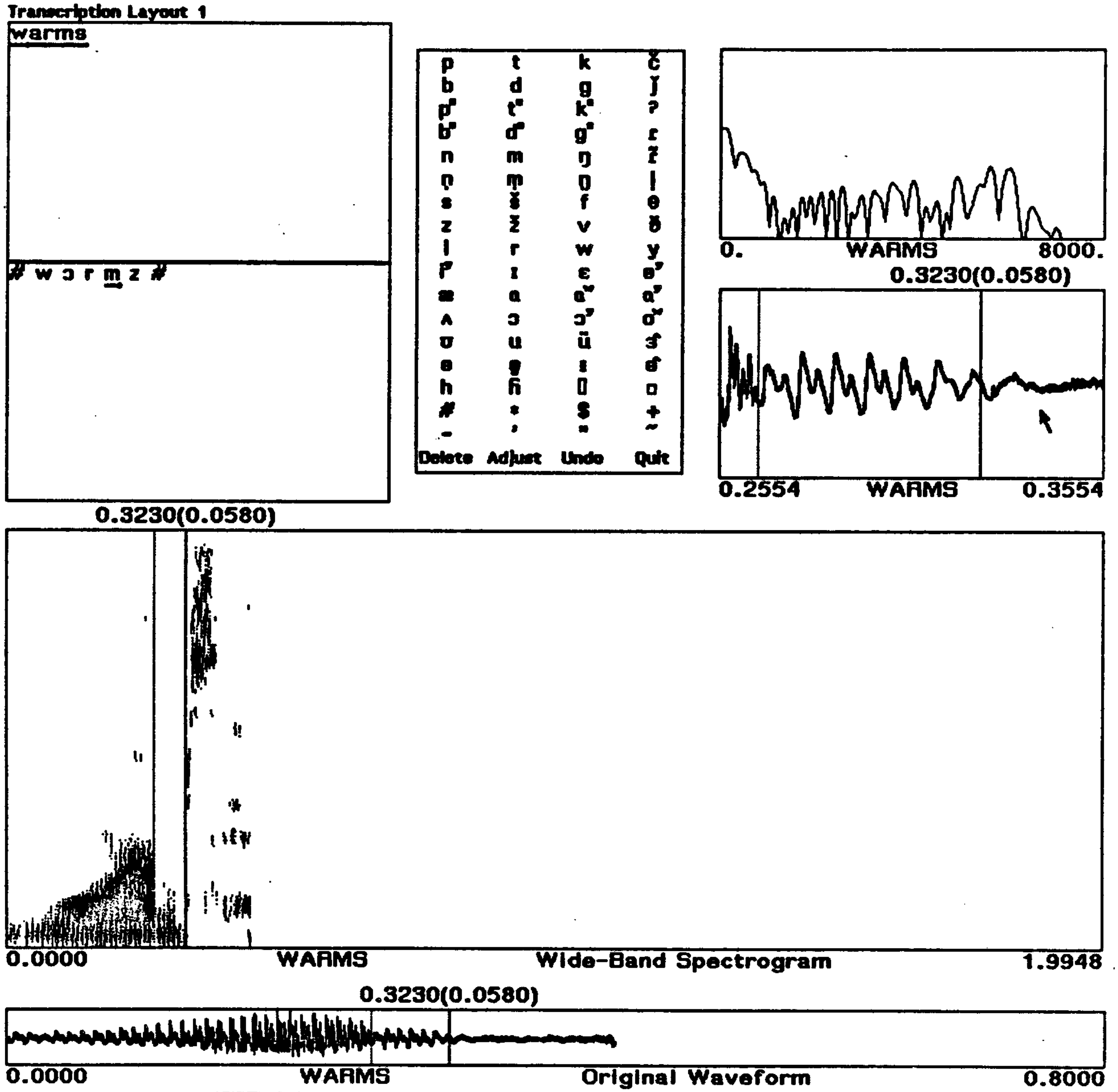
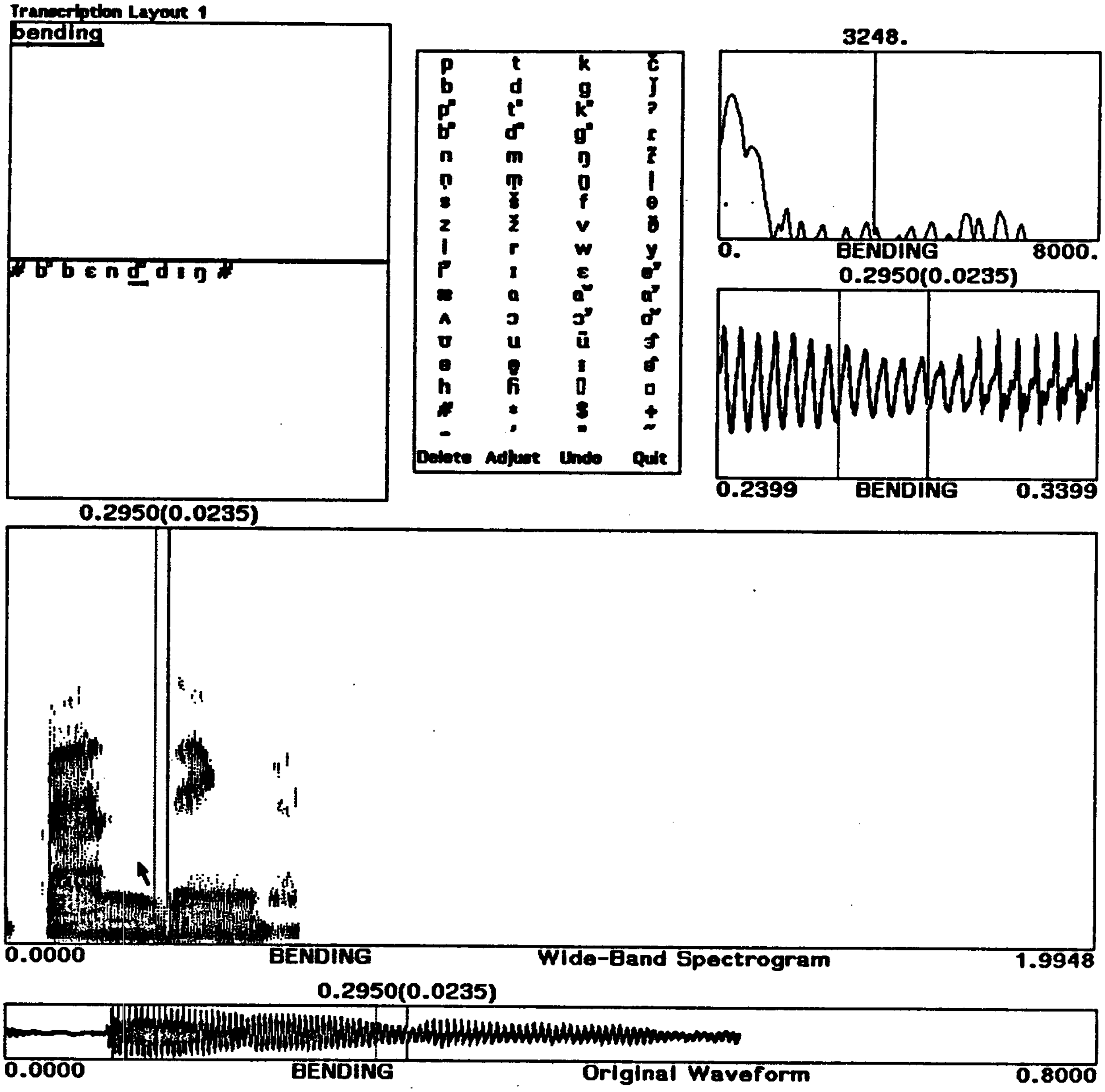


Figure B.4: The Transcription of the word *warms*



Transcription Layout 1

bending

bʰ b ɛ n d̲ʰ d ɪ ŋ

Figure B.5: The Transcription of the word *bending*

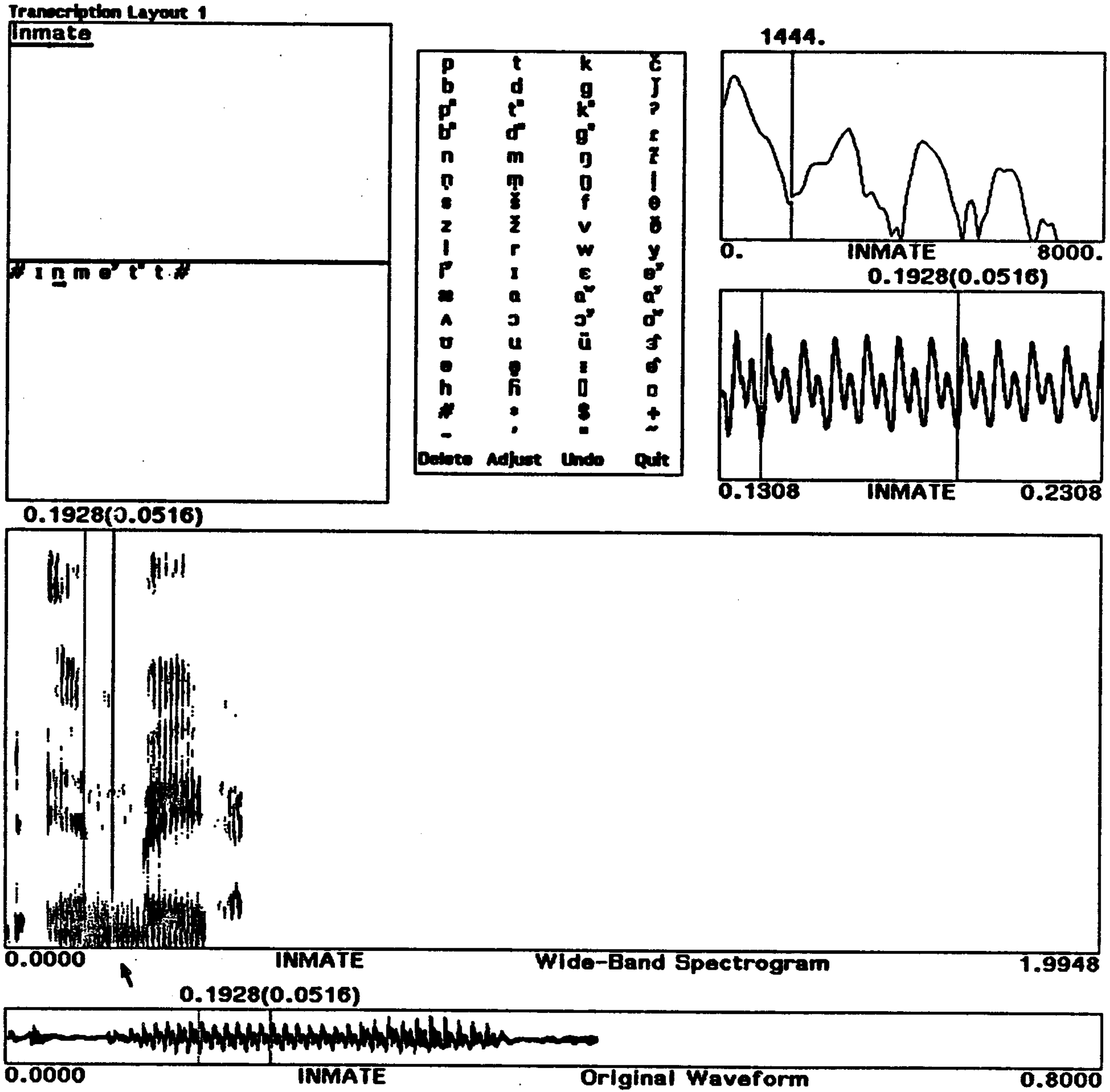


Figure B.6: The Transcription of the word *inmate*

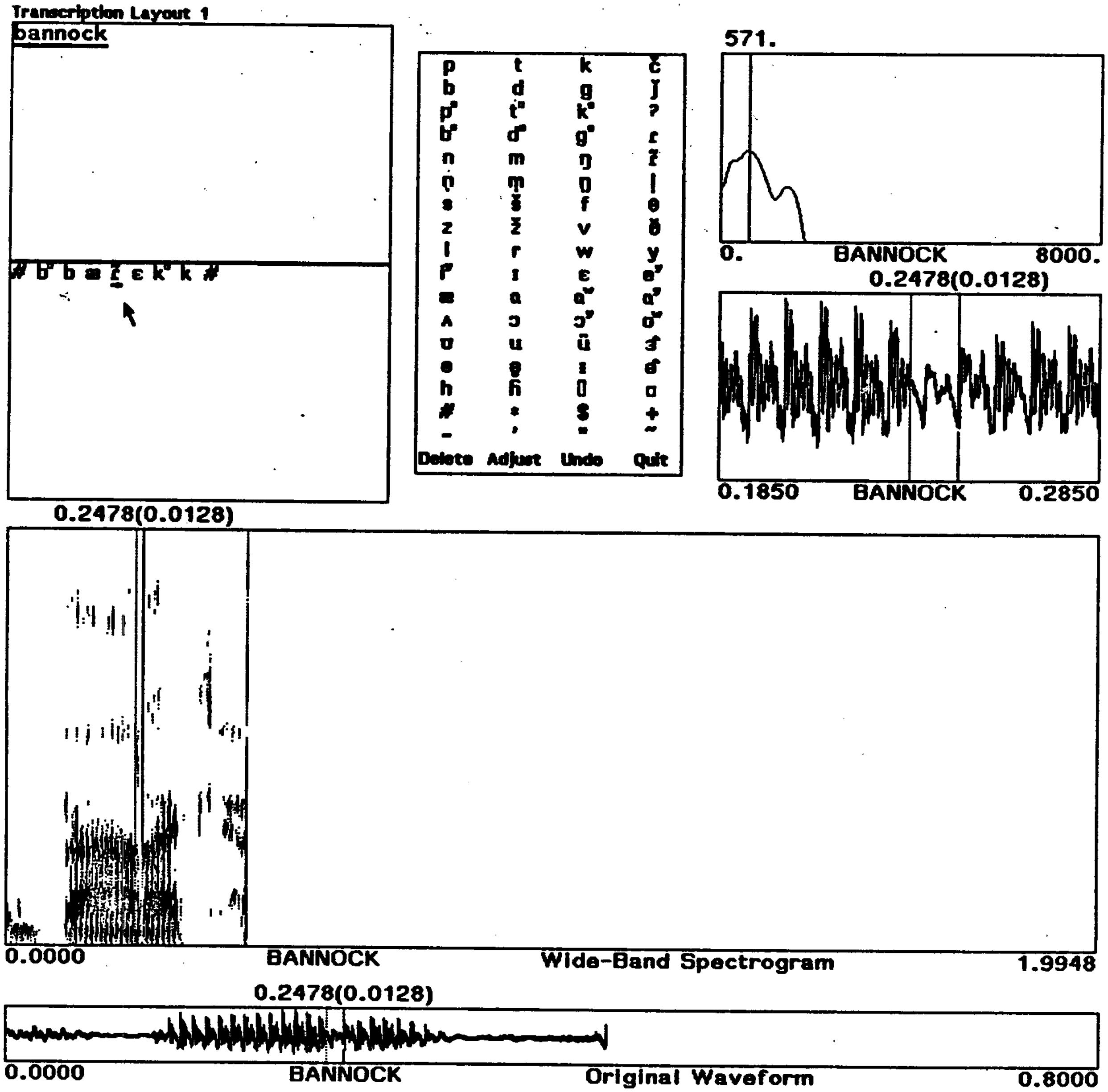


Figure B.7: The Transcription of the word *bannock*

Appendix C

Spectral Analysis Techniques

Historically, the short-time spectrum has played a major role in speech analysis. There are two main reasons for this. Acoustical studies of speech production have shown that a frequency domain representation of the speech signal succinctly captures the important acoustical characteristics of the vocal tract [13]. For example, spectral peaks in non-nasalized vowels are directly related to resonances of the vocal tract. The use of a spectral representation is also supported perceptually from clear evidence that the ear performs a form of spectral analysis at the early processing stage [14]. This indicates that acoustic features relevant to the perception of speech can be contained in a spectral representation of the speech signal. For these reasons it is desireable to compute some sort of spectral representation of the speech signal.

Conventional Fourier transforms are of little use in speech analysis since the vocal apparatus is continually changing with time. However in most speech processing schemes the vocal mechanism is considered to be quasi-stationary in that its acoustic characteristics change slowly with time [13]. This assumption motivates short-time analysis procedures in which short segments of the speech signal are isolated and processed as if they were short segments from a sustained sound (with fixed acoustic properties) [61]. Such processing produces a time-dependent sequence which in the case of the short-time spectrum reflects both the time varying nature of a speech signal and its spectral characteristics at any particular point in time.

The following sections elaborate on several different spectral analysis procedures. The intent is not to provide a rigorous mathematical background of the subject matter but to illustrate some of the issues associated with common spectral analysis techniques when they are applied to speech and, in particular, when they are applied to a statisical analysis of nasal consonants and nasalized vowels.

C.1 Short-Time Fourier Analysis

If we consider a speech signal to be represented by the time function $x[n]$, then the short-time Fourier transform is defined as

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega m} \qquad (C.1)$$

where $w[n]$ is a weighting function that determines the portion of the speech signal which receives emphasis at a particular time index, n [61]. This equation, which is a function of both time and frequency, can be interpeted in two ways. If ω is assumed to be is fixed then $X_n(e^{j\omega})$ can be considered as a form of convolution or linear filtering operation of the speech signal $x[n]$ with a filter of impulse response $w[n]e^{-j\omega n}$.[1] Typically $w[n]$ is a low-pass filter. Thus a set of outputs for different ω's and (possibly) different $w[n]$'s will result in a filter bank, commonly used for speech analysis.

A second interpretation of the short-time Fourier transform assumes that n is fixed. In this case $X_n(e^{j\omega})$ is simply the Fourier transform of the function $x[m]w[n-m]$ for $-\infty < m < \infty$. For a sampled speech signal where interest is confined to a set of equally spaced frequencies between 0 and the sampling rate F_s

$$\omega_k = \frac{F_s k}{N}, k = 0, 1, ..., N-1 \qquad (C.2)$$

[1] There are several other ways to interpret this case as well [61]. They all deal with some sort of filtering operation however.

the short-time Fourier transform can be shown to be equivalent to the discrete Fourier transform (DFT) of the windowed sequence and thus can be computed using the fast Fourier transform (FFT) algorithm [56]. This points out the fundamental similarities between a filter bank and the DFT.

Since the shape of the filter windows has a substantial effect on the output of the short-time Fourier transform, it is important to consider them carefully. The remainder of this section will focus on this issue, given that we have decided to create a filter bank for spectral analysis purposes. The type of filter bank necessary for a statistical analysis will also be discussed.

Although a number of different filter bank structures have been proposed for speech analysis, there is no simple guideline for choosing an optimal filter bank for a particular application. There are many different variables to determine including: the type of filter (IIR or FIR [61]); the filter spacing (uniform or nonuniform; overlapping or nonoverlapping); the number of filters and the filter frequency responses. One constraint which limits the range of some of these parameters is an intelligibilty requirement. The filter bank output should retain enough information of the original speech signal that it can be correctly perceived by human listeners. Thus one could evaluate the relative merit of different filter banks by performing a series of perceptual tests [68]. A similar form of judgment could be made by analyzing vocoder performances [14]. Another possible procedure to judge the merit of different filter banks would be to investigate their relative success at the front end of a standard speech recognition system [3], [8], [74]. Using this latter method, Dautrich et al. have reported a number of interesting points concerning filter banks:

- filter bank performance deteriorates for too few filters due to poor resolution.
- filter bank performance deteriorates for too many nonoverlapping filters since their frequency responses become so narrow that some end up measuring noise between pitch harmonics.

- successful filter banks have an essentially flat overall frequency response without sharp peaks or valleys.

- the best performance of non-uniform filter banks is obtained for filters spaced along a critical band frequency scale as opposed to octave bands, $\frac{1}{3}$ octave bands, or arbitrary spacing.

Of course, it is possible to design a filter bank solely on psychophysical data. The motivation for this is that a spectral analysis would then emphasize the things which are known to be perceptually important and demphasize those which are not. For instance, a small change in the frequency of a higher formant should not be as important as the same change in the first formant because the just-noticable difference (JND) for a formant frequency increases with frequency [29]. Several attempts have been made to design filter banks with these considerations in mind [67],[80].

What kind of filter bank could be used for a statistical analysis of speech sounds? Optimally, the exact filter bank shape should not be critical to the success of an analysis experiment. Global spectral features should be able to be established by the data independent of any particular filter bank shape. Specific filter bank details could be dictated somewhat by the particular kind of spectral analysis being performed. For an analysis of the steady state portion of nasal murmurs or nasalized vowels for instance, a long filter impulse response or time window could be used since there are no sharp temporal changes in these regions. The main advantage of using a long filter window is that one can obtain good spectral stability independent of the window position relative to the pitch period as illustrated in figures C.1 – C.3 for a synthetic steady-state vowel with a fundamental frequency of around 100 Hz.

In figure C.1 we see that two hamming windows (with 7 ms and 25 ms duration) have been centered at the beginning of a pitch period. Both of their corresponding DFT's show a good spectral representation of the vocal tract. Note that the pitch harmonics are visible in the DFT of the 25 msec hamming window because of the tradefoff between time and frequency resolution. Figure C.2 illustrates the same conditions except that the windows have been centered at the tail end of the preceeding pitch period. As one would expect, the DFT of the shorter hamming window yields a very poor spectral representation of the resonances of vocal tract. At this point in the pitch period, the glottal folds are open so that the resonances of the vocal tract are severely damped. Figure C.2 also shows that the longer hamming window is able to extract a reliable vocal tract shape since it overlaps multiple pitch periods. Figure C.3 illustrates the effect of centering the windows in the middle of a pitch period. Clearly a longer window length will yield a more stable spectral response.[2]

The penalty for spectral stability is reduced temporal resolution as is illustrated in figures C.4 – C.6 for another synthetic vowel. In this example the first and third resonances are held fixed at 450 and 2450 Hz respectively while the second resonance is changed from 950 to 1950 Hz within 10 msec. In figure C.4 the windows are centered at the start of the last pitch period before the start of the transition period. Note that the DFT of the longer hamming window reflects the fact that the window overlaps multiple pitch periods with different vocal tract characteristics since the second resonance range is smeared in the spectral domain. This holds true for figures C.5 and C.6 as well. Notice that in each case the shorter hamming window isolates a single pitch period which produces a superior spectral shape.

Clearly if we knew the exact location of the pitch period, it would be possible to center a window over the important part of the pitch period to produce an excellent estimate of the vocal tract spectral shape. The duration of the window could be chosen to be short enough so that there would be no pitch information present in the spectral domain resulting in a smooth spectral shape. Further, the spectra would be stable with time yet accurately reflect changes in the vocal tract.

[2] A 25 msec window starts to show instability if the pitch period is much below 12.5 msec (80 Hz).

Unfortunately, automatic pitch-synchronous analysis is difficult to perform reliably. Most analysis procedures locate the pitch period boundaries either manually or at most semi-automatically [47], [58], [65]. In any event, automatic pitch-synchronous analysis is beyond the scope of this thesis. As was illustrated in the previous figures, a long window is probably the best choice for asynchronous analysis of steady state sounds.

For spectral analysis of nasal consonants and nasalized vowels it seems reasonable to use a long window which would provide good spectral resolution. Temporal resolution is not a critical factor here because the analysis will take place in a relatively stationary environment.

As a first pass at analysis, a uniformly spaced filter bank was used. A hamming window was chosen because of its superior spectral properties.[3] For spectral stability for most pitch frequencies a long duration window (25 msec) was used.

From previous figures it is clear that with decreased time resolution one obtains superior spectral resolution. However this is detrimental to any study of spectral shapes since one does not want pitch information in the spectral estimate of the vocal tract. Thus some form of spectral smoothing is necessary. One solution to this problem would be to smooth the cepstrum of the speech signal.[4] Cepstral smoothing or homomorphic filtering has been successfully used for many speech processing applications where smoothed spectra are necessary [52], [55], [66]. Figures C.7 – C.9 show a plot of the cepstrum, the cepstrum window and the unsmoothed and smoothed FFT spectra for various pitch values, ranging from 120 Hz to 440 Hz, on a synthetic vowel. For the purposes of this analysis the smoothing window implemented was 3 msec long (flat for 1.5 msec and tapered with a raised cosine for 1.5 msec). This window was found to produce acceptably smooth spectra for pitch frequencies below 300 Hz. This is acceptable for most

[3] This data window is attractive because the side lobes of its Fourier transform remain more that 40 dB down at all frequencies [56].

[4] The cepstrum is defined as the Inverse Fourier transform of the log magnitude spectrum [55].

male and female speakers but not for pitch frequencies of some children. The problem is avoided since the speech of children is not is not analysed in this study.

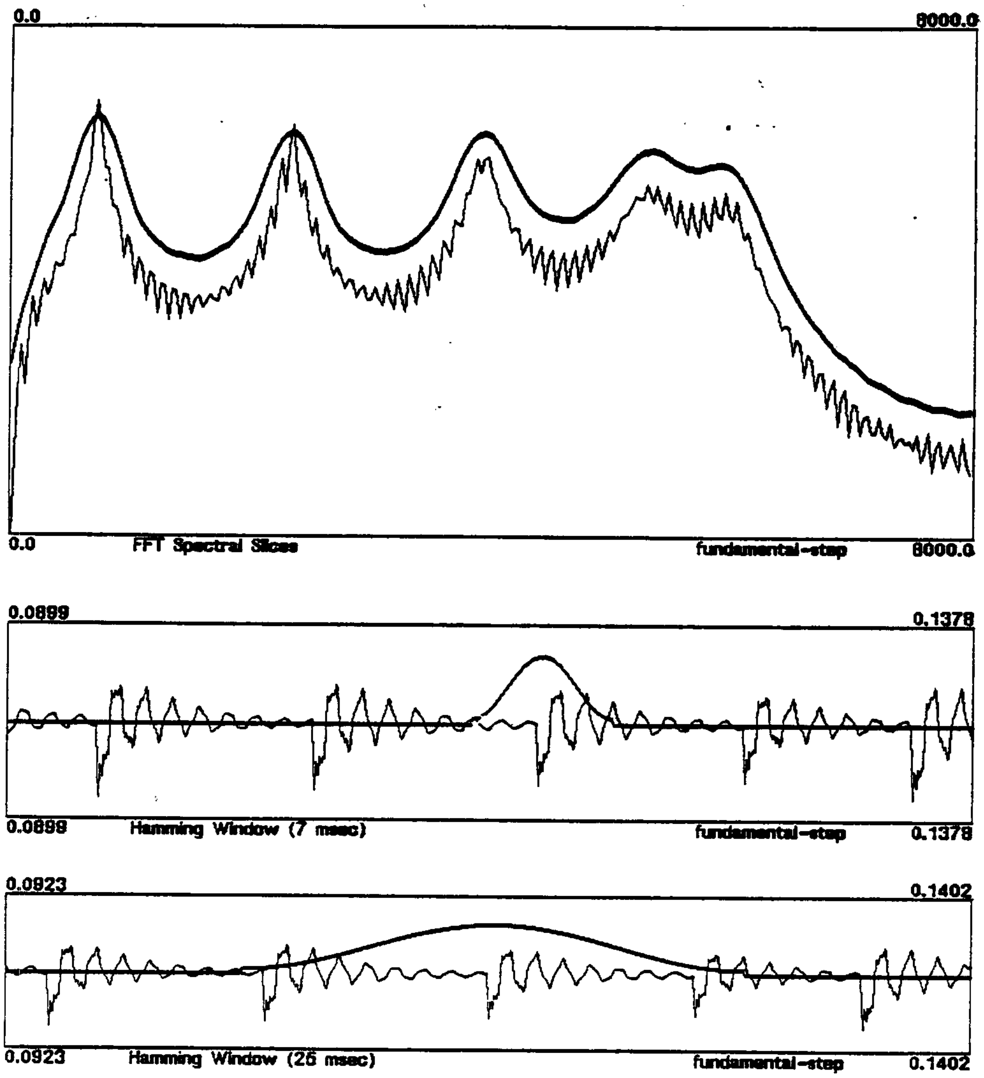


Figure C.1: Hamming Windows Centered at Start of Pitch Pulse

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

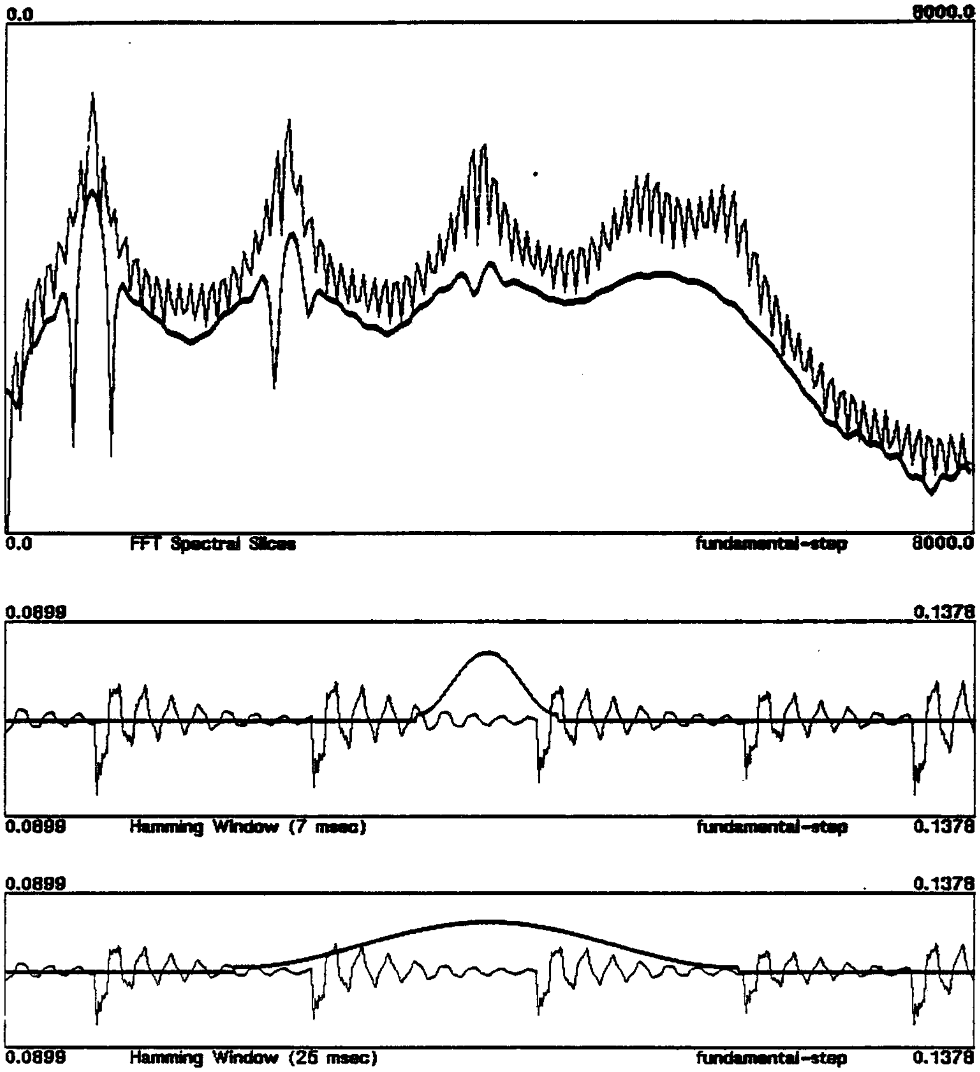


Figure C.2: Hamming Windows Centered at End of Pitch Pulse

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

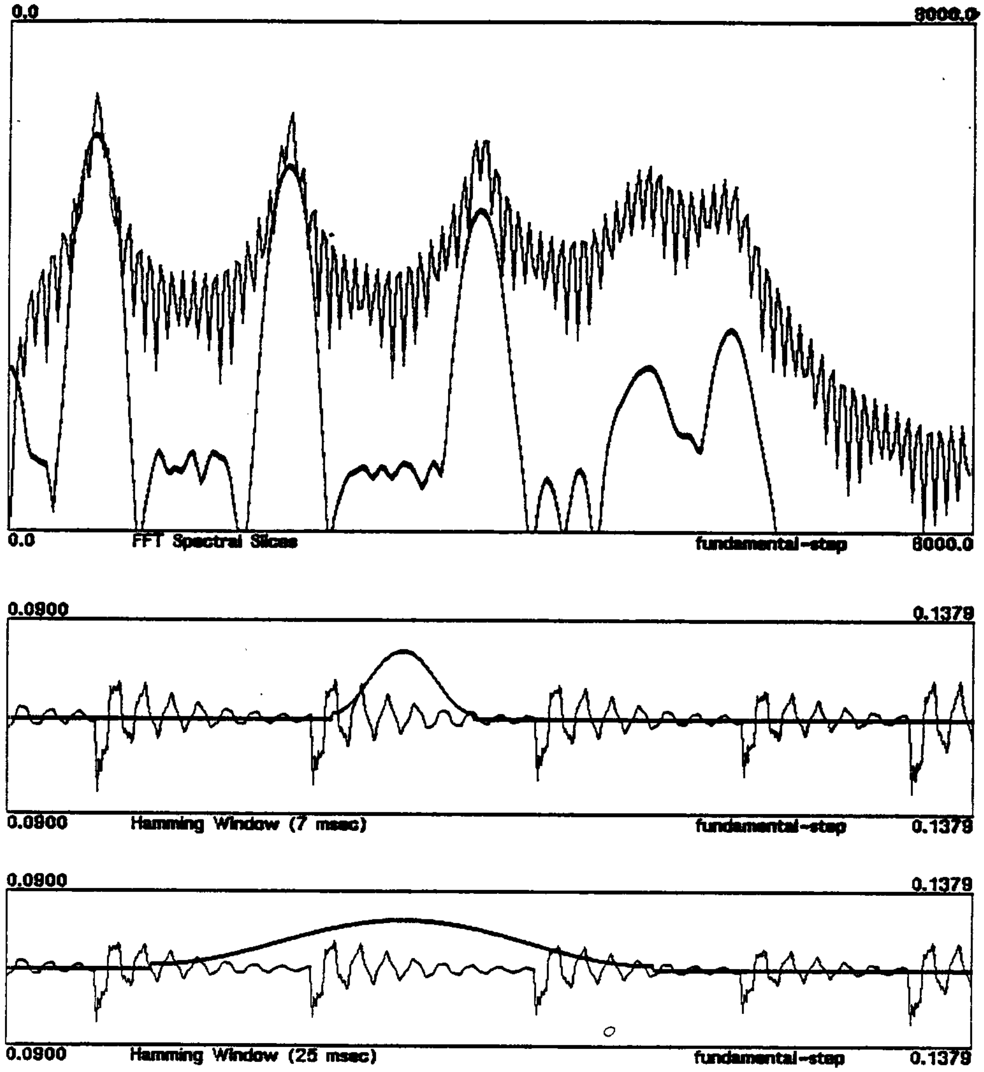


Figure C.3: Hamming Windows Centered at Middle of Pitch Pulse

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

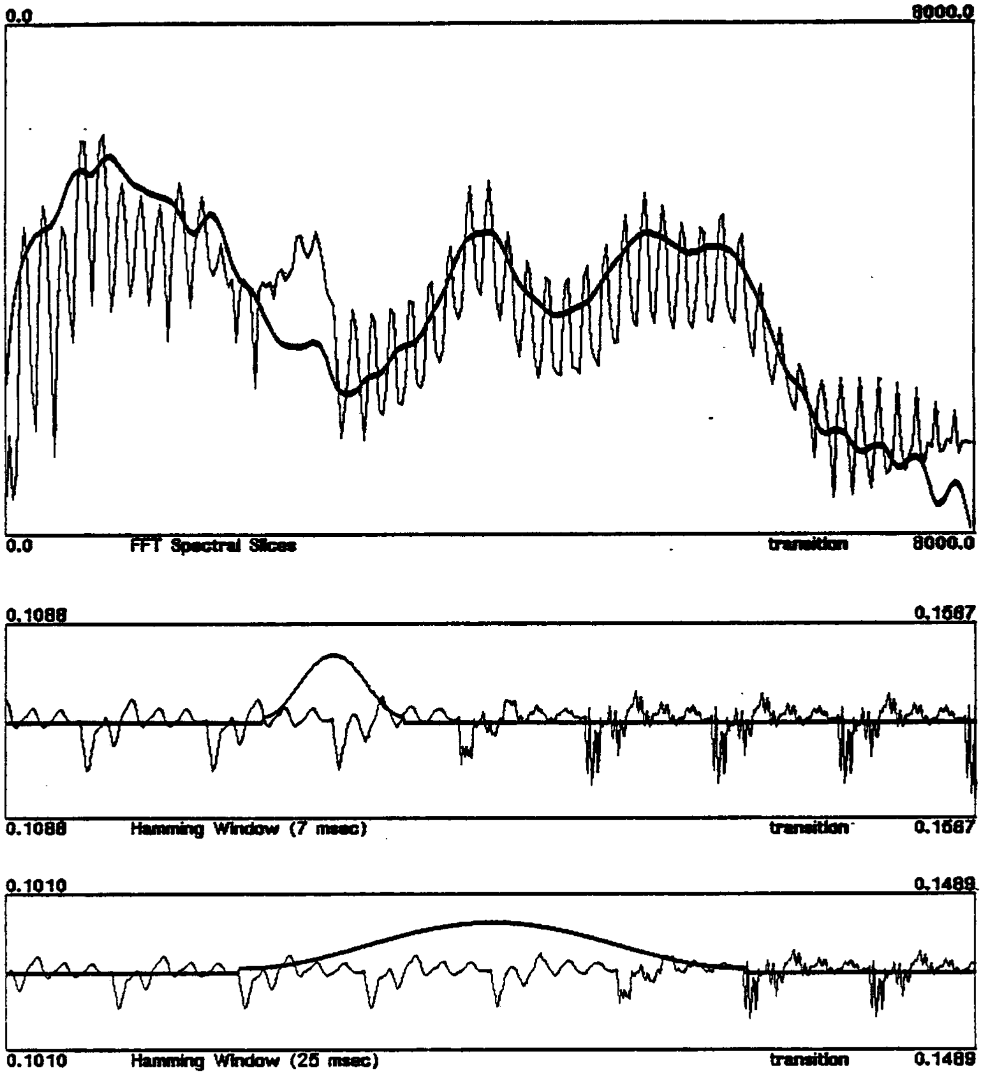


Figure C.4: Hamming Windows Centered at Start of Formant Transition

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

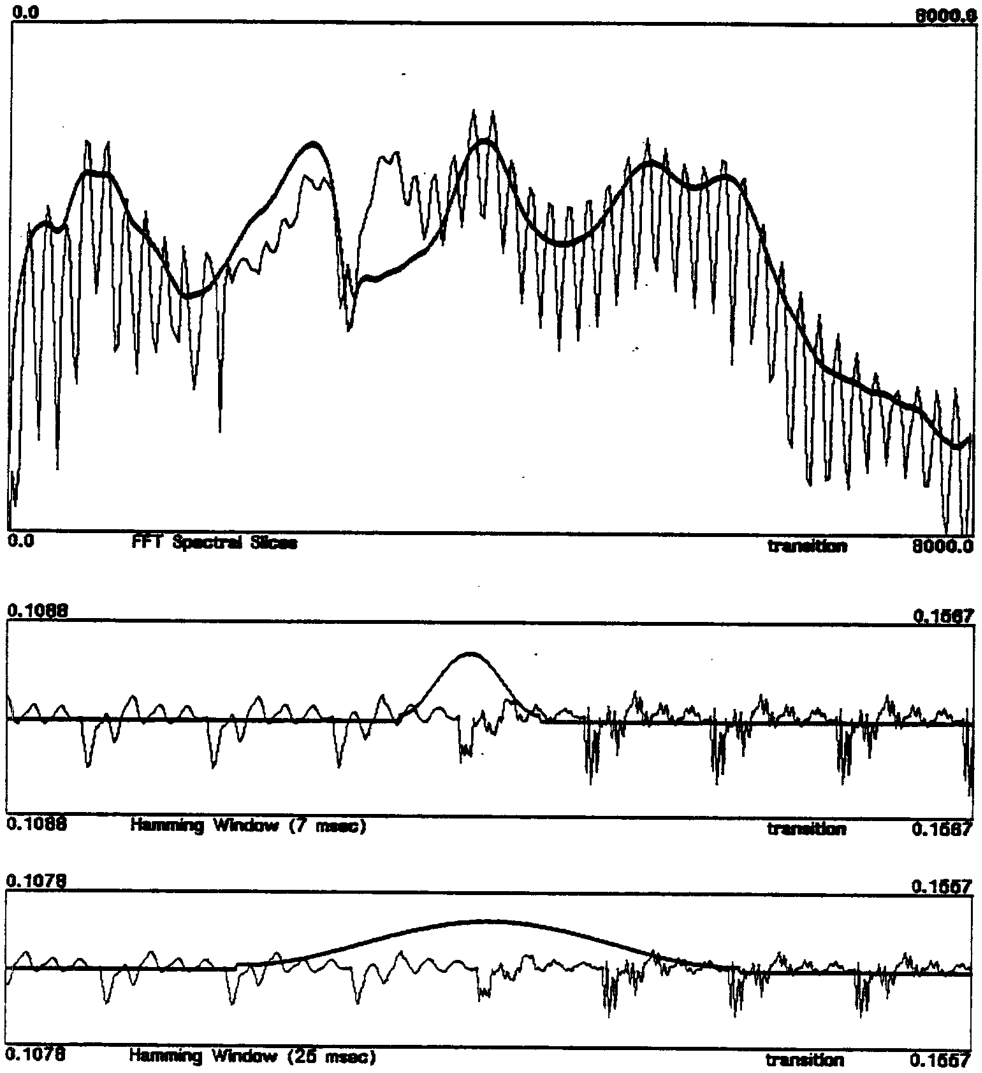


Figure C.5: Hamming Windows Centered at Middle of Formant Transition

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

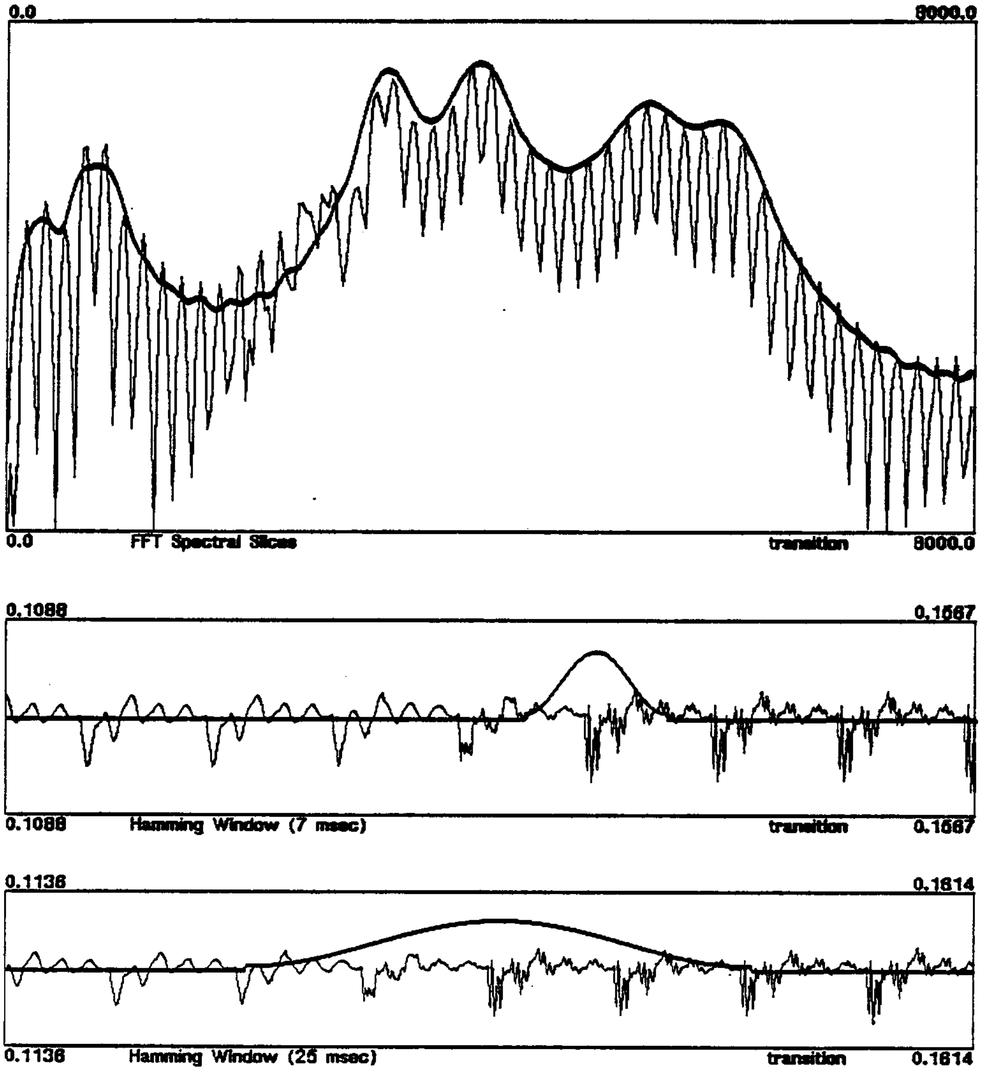


Figure C.6: Hamming Windows Centered at End of Formant Transition

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

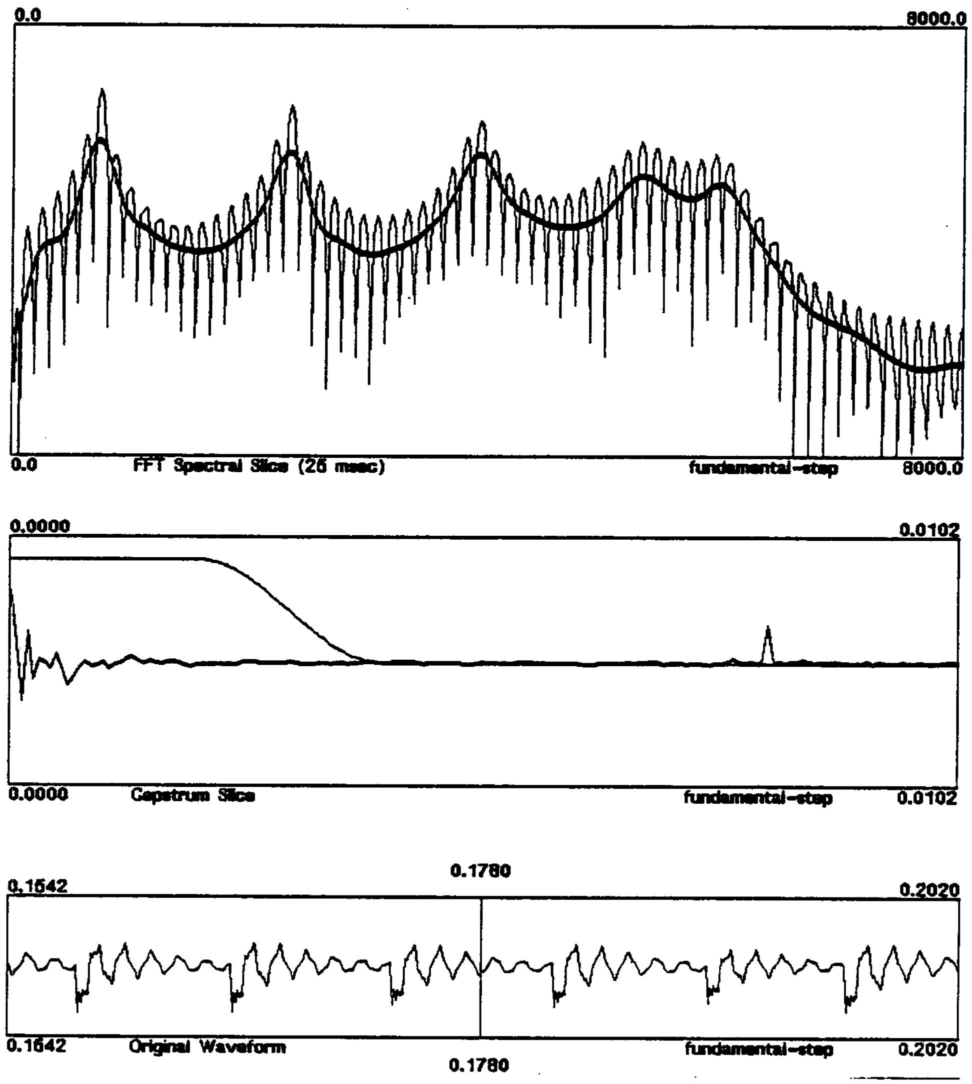


Figure C.7: Cepstrally Smoothed Spectra with Low Pitch Frequency

The top display contains the outputs of unsmoothed and smoothed DFT spectra (hamming window 25 msec). The middle display contains the cepstrum and the smoothing window (2 msec flat, 2 msec raised cosine). The bottom display shows the original speech waveform (pitch approximately 120 Hz).

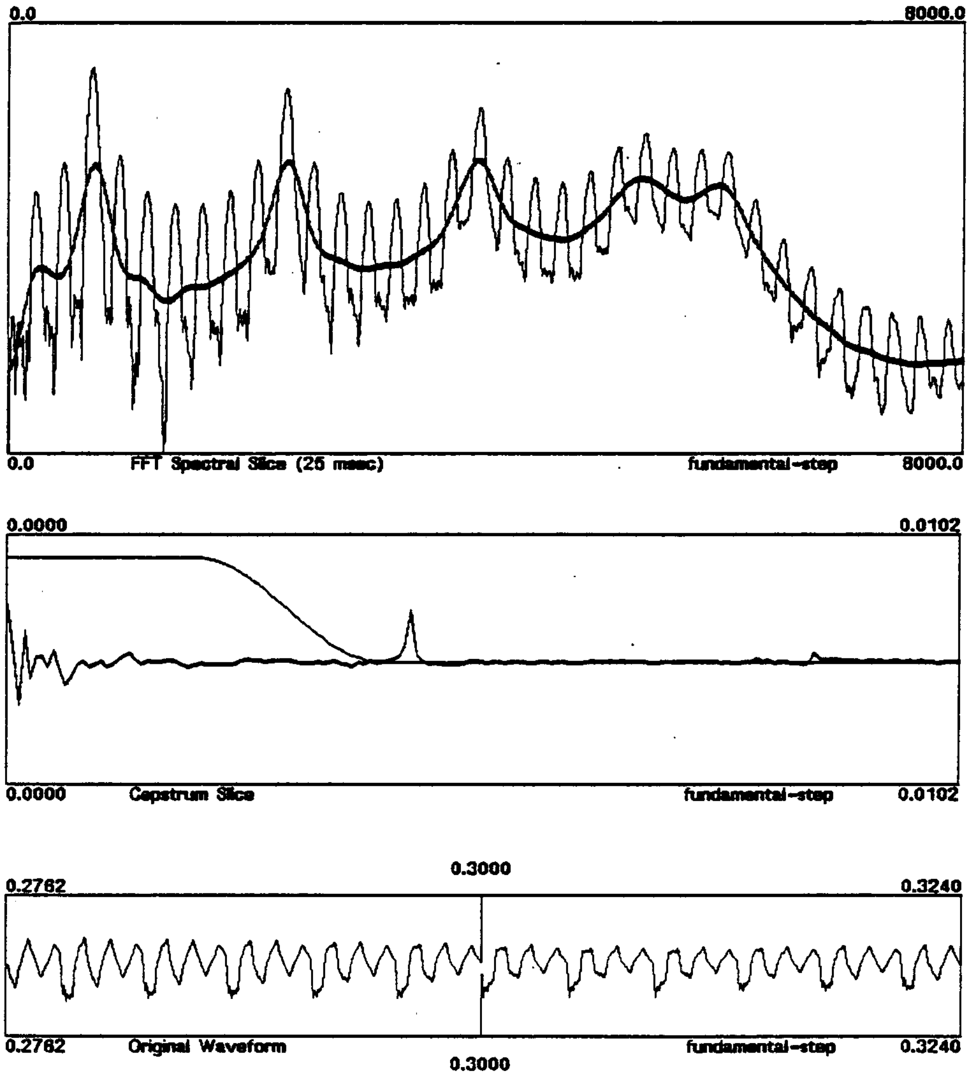


Figure C.8: Cepstrally Smoothed Spectra with Middle Pitch Frequency

The top display contains the outputs of unsmoothed and smoothed DFT spectra (hamming window 25 msec). The middle display contains the cepstrum and the smoothing window (2 msec flat, 2 msec raised cosine). The bottom display shows the original speech waveform (pitch approximately 220 Hz).

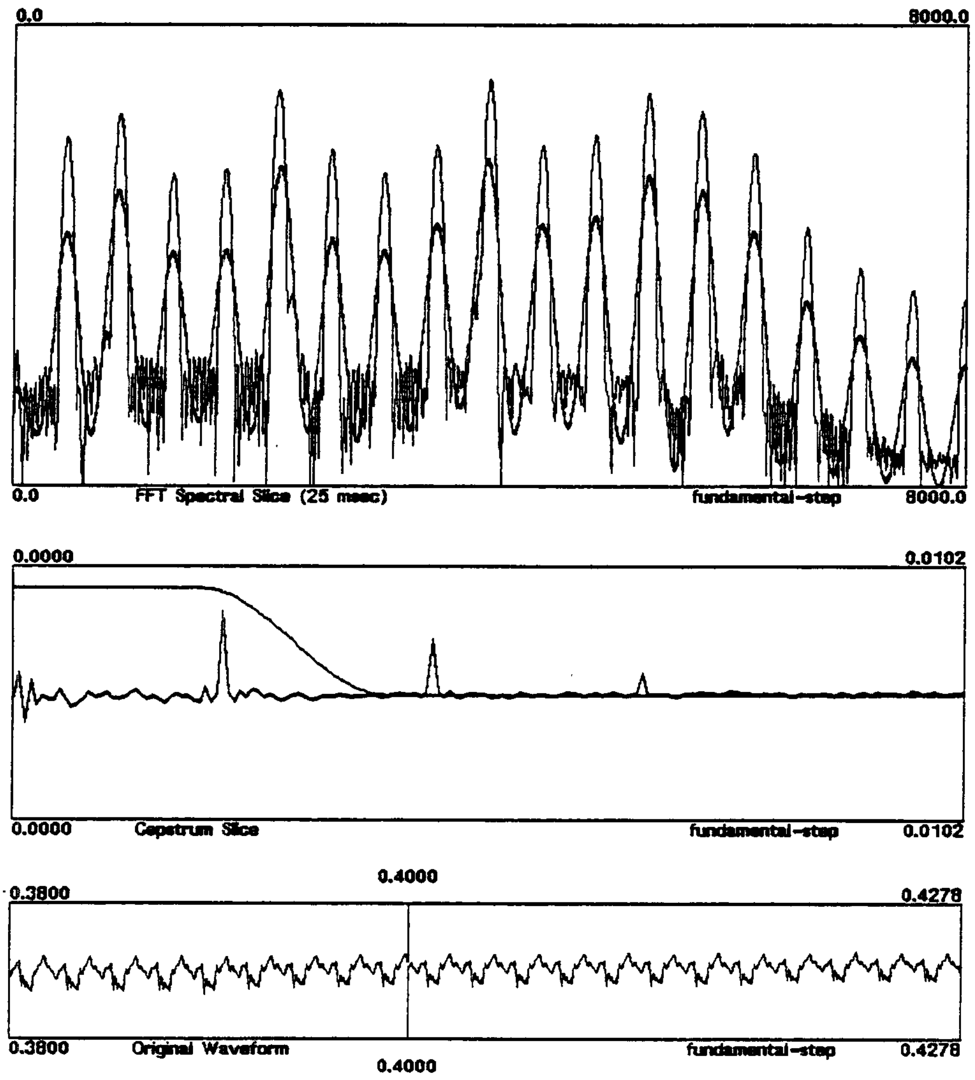


Figure C.9: Cepstrally Smoothed Spectra with High Pitch Frequency

The top display contains the outputs of unsmoothed and smoothed DFT spectra (hamming window 25 msec). The middle display contains the cepstrum and the smoothing window (2 msec flat, 2 msec raised cosine). The bottom display shows the original speech waveform (pitch approximately 400 Hz).

C.2 Vocal Tract Modelling

An alternative to spectral representations based on filter banks or homorphic analysis is to use an approach based on estimating the parameters of a vocal tract model. In fact for most models the vocal tract response $V(z)$ is considered as only one part of the overall frequency response of the speech signal $H(z)$. In general the glottal pulse and radiation components, $G(z)$ and $R(s)$ are taken into account as well.

$$H(z) = G(z)V(z)R(z) \tag{$C.3$}$$

For this model, the original input is considered to be a train of impulses at the pitch period.

One such model could consist of representing the overall transfer function in terms of a general transfer function of the form

$$H(z) = G\,\frac{\prod_{i=1}^{q}(z - z_i)}{\prod_{i=1}^{p}(z - z_i)} \tag{$C.4$}$$

where the parameters used to describe the speech signal are the poles and zeros of the transfer function and the gain factor G. In general, the impulse response associated with the transfer function is a nonlinear function of the numerator and denominator coefficients. Estimating these parameters for a segment of speech would thus typically require the solution of a set of nonlinear equations. For the special case in which the order of the denominator polynomial is zero, the determination of the parameters based on a mean-square error criterion reduces to the solution of a set of linear equations. For the case where the order of the numerator polynomial is zero, the mean-square error criterion reduces to the solution of a set of linear equations of the inverse filter. All-pole modelling is very common for speech analysis and is commonly known as Linear Predictive Coding (LPC) [42], [45].

One important attribute of the vocal tract transfer function is that it is characterized primarily by resonances which are well represented by poles. However, difficulties can arise when the model is invalid, as is true for nasal consonants and nasalized vowels. Figure C.10 shows examples of synthetic stimuli with zeros included at low frequencies. In the top display the zero is located at 1000 Hz as might be found in a nasal consonant. This zero creates a dip in the DFT spectra which is not captured by the LPC spectra. In the bottom display the zero is located at 450 Hz between two poles as might be found in a nasalized vowel. This pole-zero-pole combination is again not captured in the LPC representation although it exists in the DFT spectra. These figures can be compared to cepstrally smoothed spectra as shown in figure C.11 where the essence of the DFT spectra have been captured satisfactorally.

Clearly it is possible to modify the model so that one is better able to match the DFT spectra. For instance in the above examples it is possible to use more poles (19 poles were used for 8000 Hz bandwidth), or to attempt pole-zero modelling. However, no modelling procedure will work correctly all of the time. For this reason it was decided to use a spectral representation such as cepstrally smoothed spectra which does not rely on any underlying model of the speech waveform, and so will tend to be more robust.

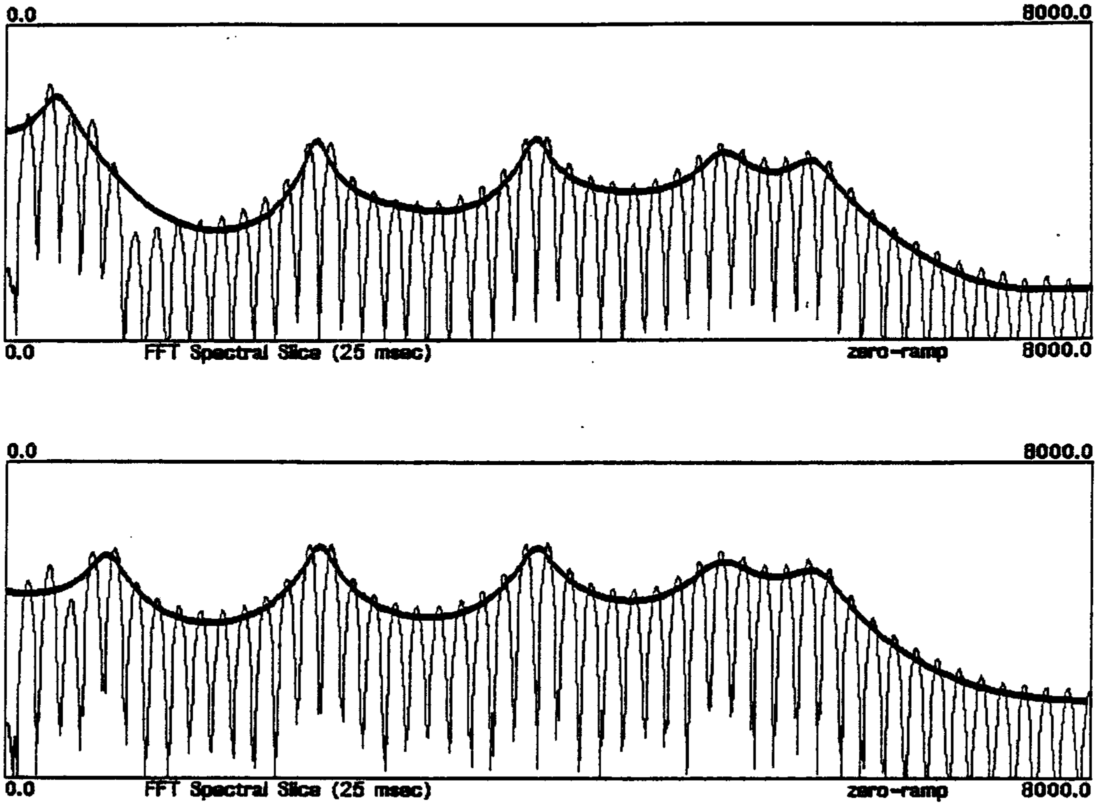


Figure C.10: LPC Spectra

The top display contains the outputs of DFT and LPC spectra (dark line) of a synthetic token containing a zero at 1000 Hz. The bottom display contains the outputs of DFT and LPC spectra (dark line) of a synthetic token containing a zero at 450 Hz. Hamming windows of 25 msec duration were used.

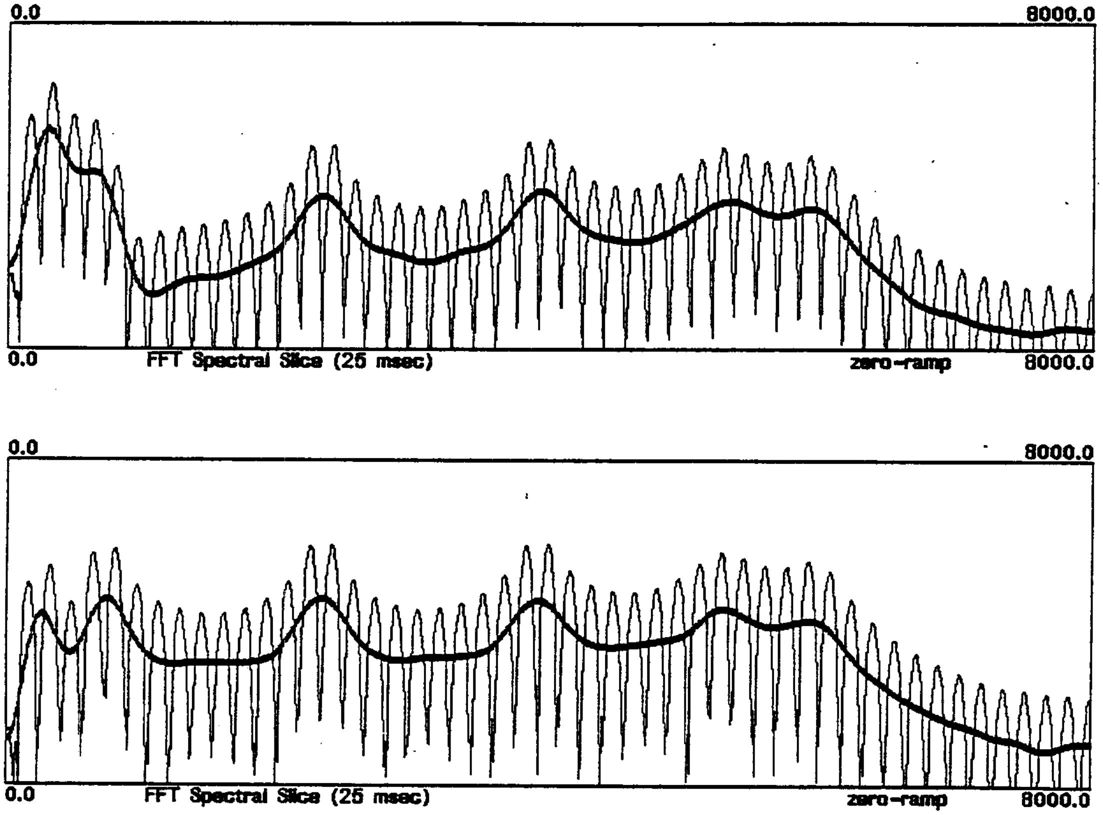


Figure C.11: Cepstrally Smoothed Spectra

The top display contains the outputs of DFT and cepstrally smoothed spectra (dark line) of a synthetic token containing a zero at 1000 Hz. The bottom display contains the outputs of DFT and cepstrally smoothed spectra (dark line) of a synthetic token containing a zero at 450 Hz. Hamming windows of 25 msec duration were used.

Appendix D

Nasalized Vowel Algorithms

Since nasality manifested itself more subtly in nasalized vowels than in nasal consonants, the algorithms used to extract this property from vowels were more sophisticated, and as a result, more fragile, than those used for the nasal consonants. There were two types of calculations used for the analysis of nasalized vowels: those which performed general statistical measures of the spectral distributions, and those which performed peak picking, and measured properties of actual resonances. Both types of computations were found to be effective in distinguishing nasalized vowels from non-nasalized vowels. The following sections describe the algorithms used for each type of calculation.

D.1 Statistical Calculations

Center of Mass

Since the center of mass was found to be a rather ineffective measure of nasalization, its main role was to find the center of energy in the low frequency regions so that the local spread of energy could be measured by the standard deviation calculation.

The center of mass is a specific kind of spectral weighting algorithm, described in chapter 2, where the weighting window, $\vec{W}$, is linear with frequency. Calculated between two frequency ranges, f_1, and f_2, the center of mass, $\bar{f}$, is defined as

$$\bar{f} = \frac{1}{A_1} \sum_{f=f_1}^{f_2} f X(f) \qquad (D.1)$$

$$A_1 = \sum_{f=f_1}^{f_2} X(f) \qquad (D.2)$$

where $X(f)$, is the value of the DFT spectra at frequency f. For use in nasalized vowels, the center of mass was computed between 0 and 1000 Hz, which covers the first formant range of most men and women [59].

In order to reduce the sensitivity of the center of mass function to sudden changes at the end points, such as a formant passing below 1000 Hz, the DFT spectra was windowed with a trapezoidal window before the center of mass was computed. The window was flat between 100 Hz and 900 Hz, and had 100 Hz tapers at each end. Windowing the spectra ensured that there were no sudden changes in the center of mass caused by a marginal movement in energy across the upper boundary.

There are several different spectral representations on which the center of mass could have been computed (magnitude squared, or magnitude for instance). However, the log magnitude squared (dB) spectrum was used because it was observed that the extra resonance frequency had the largest effect on the center of mass in this representation. In any other representation, the major resonance peak dominated the value of the center of mass.

Using the log spectrum introduced a sensitivity problem into the calculation however. In a magnitude spectra the baseline value for the center of mass is zero. There is no such corresponding baseline value for the dB scale however since values may go to $-\infty$. Thus, some form of normalization is necessary. Typical normalization procedures establish some baseline value, relative to a value in the spectrum. Note that the center of mass may be made arbitrarily sensitive this way. In this research, a good value was found to be somewhere around 20 dB below the spectral peak in the frequency range of interest. This yielded a center of mass which was responsive to changes in the first formant frequency and nasality in the vowel, but was not overly sensitive to minute changes in the spectrum.

Standard Deviation

A measure of the local spread of energy around the center of mass was found to be a very good measure of nasalization. This was calculated by measuring the second moment of local energy around the center of mass. The term "local" was defined to include all energy within a specified frequency radius of the center of mass. Thus, if the center of mass was measured to be 700 Hz, and the frequency radius, f_r, had been defined as 200 Hz, then the standard deviation would be calculated between 500 and 900 Hz. In general, the standard deviation, σ, is calculated between $\bar{f} - f_r$, and $\bar{f} + f_r$, and is defined as

$$\sigma = \sqrt{\frac{1}{A_2} \sum_{f=\bar{f}-f_r}^{\bar{f}+f_r} X(f)(f-\bar{f})^2} \qquad (D.3)$$

$$A_2 = \sum_{f=\bar{f}-f_r}^{\bar{f}+f_r} X(f) \qquad (D.4)$$

The same issues which were discussed for center of mass apply here. Thus the standard deviation was computed on the same normalized log magnitude spectrum which was used to calculate the center of mass.

The frequency range is a very important parameter since it determines the type of deviation that is being measured. The most effective range was found to be 500 Hz on either side of the center of mass. Thus, the standard deviation was measuring the overall spread of energy in the low frequency region, rather than the local spread of the first formant.

Since the center of mass was measured between the ranges of 0 to 1000 Hz, at least one end point in the standard deviation calculation would extend outside the center of mass endpoints (unless the center of mass was exactly 500 Hz). In order to include energy outside the first formant region, which was detrimental to the standard deviation measure, the standard deviation only in the valid regions. In other words, if the center of mass was 700 Hz, the standard deviation was computed between 200 Hz and 1000 Hz.

Although the frequency range restriction was necessary, it made the value of the standard deviation measure frequency dependent. In fact, the maximum value of the standard deviation at any frequency, would be linearly related to the width of the frequency region used in the calculation. Thus, if a deviation value was computed over an 800 Hz range, its maximum value could only be 0.8 that of a deviation which used a full 1000 Hz range. In an attempt to normalize the standard deviation, so that it was frequency independent, each value was scaled upwards by the ratio of the maximum frequency width (1000) to the actual frequency width used in the calculation. This procedure was found to substantially reduce the frequency dependence of the standard deviation calculation.

D.2 Resonance Calculations

Qualitative observations indicated that it would be useful to measure certain properties of the actual resonances in the low frequency region of the spectra. Before this could be done , the resonances themselves had to be found. To do this, spectral regions were established by searching for zero crossings in the second derivative of the smoothed log spectra, as was illustrated in chapter 2. Once the spectral regions were established, resonances could be found by collecting all the peak regions below 1100 Hz.

The actual collection algorithm only gathered resonances until it either had two, since one would be a first resonance and the other a nasal resonance, or until it had passed over 1000 Hz. Note that this procedure introduces a flaw into the system, since non-nasalized high back vowels could be collected if the second formant was below 1000 Hz. The magnitude of this problem was reduced by checking to make sure that if there were two peaks collected, one of them was actually below 400 Hz. This ensured that at least one resonance was either a nasal resonance, or a very low first formant. Thus, if two resonances were found at 500 Hz and 800 Hz, the 800 Hz resonance would be rejected, and the 500 Hz resonance would be kept. Thus, the main vowel which caused problems with this sorting algorithm was /u/.

Percentage

Once the two lowest resonances were established, the percentage measure was calculated in a given time region by dividing the number of spectral slices which had two resonances in the time region, by the number of spectral slices in the time region.

Resonance Dip

Whenever there were two resonances in the spectrum, the resonance dip was calculated by measuring the difference, in dB, between the smallest resonance, and the valley. If there were not two resonances. no value of resonance dip was computed.

Resonance Difference

Whenever there were two resonances in the spectrum, the resonance difference was calculated by measuring the difference, in dB, between the second resonance, and the first resonance. Note that no attempt was made to determine which resonance was the nasal resonance. Thus for high vowels, the resonance difference was negative, and for low vowels, the resonance difference was positive. If there were not two resonances, no value of resonance difference was computed.