AUTOMATED NASAL FEATURE DETECTION FOR THE LEXICAL ACCESS FROM FEATURES PROJECT

by

Neira Hajro

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

Massachusetts Institute of Technology

April 14, 2004
[June 2004]

Author _____
Department of Electrical Engineering and Computer Science
April 14, 2004

Certified by _____
Kenneth N. Stevens
Thesis Supervisor

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

AUTOMATED NASAL FEATURE DETECTION FOR THE LEXICAL ACCESS FROM FEATURES PROJECT

by

Neira Hajro

Submitted to the Department of Electrical Engineering and Computer Science

April 14, 2004

In Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

The focus of this thesis was the design, implementation, and evaluation of a set of automated algorithms to detect nasal consonants from the speech waveform in a distinctive feature-based speech recognition system. The study used a VCV database of over 450 utterances recorded from three speakers, two male and one female. The first stage of processing for each speech waveform included automated `pivot` estimation using the Consonant Landmark Detector – these `pivots` were considered possible sonorant closures and releases in further analyses. Estimated pivots were analyzed acoustically for the nasal murmur and vowel-nasal boundary characteristics. For nasal murmur, the analyzed cues included observing the presence of a low frequency resonance in the short-time spectra, stability in the signal energy, and characteristic spectral tilt. The acoustic cues for the nasal boundary measured the change in the energy of the first harmonic and the net energy change of the 0-350Hz and 350-1000Hz frequency bands around the pivot time. The results of the acoustic analyses were translated into a simple set of general acoustic criteria that detected 98% of true nasal pivots. The high detection rate was partially offset by a relatively large number of false positives – 16% of all non-nasal pivots were also detected as showing characteristics of the nasal murmur and nasal boundary. The advantage of the presented algorithms is in their consistency and accuracy across users and contexts, and unlimited applicability to spontaneous speech.

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Dedication:

To Dad, Mom, and Lejla

For their love and support

Posvećeno:

Tati, Mami i Lejli

Za svu ljubav i podršku

Acknowledgments

First I would like to thank my thesis advisor, Kenneth Stevens, for giving me the opportunity to work in the Speech Communications laboratory. His guidance, support, and help have been crucial in the completion of this thesis and my experience as a graduate student at MIT.

I would like to thank Janet Slifka for keeping her office door open for questions, concerns, and problems throughout my time in the lab. Her help with MATLAB was invaluable throughout my research work. My gratitude also goes to Tony Okobi and Steven Lulich for entertaining various questions and discussions regarding the results of my data analyses, and patience in sharing their extensive knowledge. My thanks also go to Vergilio Villacorta, Sherry Zhao, and the rest of the Speech Group for many fun afternoons spent in the lab.

I could not imagine a better academic advisor than Prof. Patrick H. Winston – I am indebted for having him as my advisor, teacher, and most of all, a friend. Without Anne Hunter and Vera Sayzew I would not have found my place in the EECS department or at MIT. I am also grateful to my friends and many social breaks we entertained together – I am especially thankful to Rados Radoicic, Ivana Kalea, and Vedran Peric for their support and friendship during the writing of this thesis.

I would like to express my gratitude to my host family in Virginia – without their help I would not be at MIT today.

I am indebted to Noshirwan Petigara for his encouragement, support, and advice throughout my years at MIT.

Most of all, I would like to thank my parents. Their love, support, and acceptance gave me the ambition and strength to succeed. I am indebted to my sister Lejla for growing with me both professionally and personally, so that I have company every step of the way.

Chapter 1

1.1 Introduction

A multitude of speech recognition models have been developed over the past two decades. Many of these models use general pattern matching techniques, with little or no speech specific knowledge. In the pattern-matching approach, a model contains a training set and an operating set of words. The training set contains a selection of words from the lexicon and is adapted over time to more closely resemble the speaker's pronunciation. The operating set includes a set of words spoken by the speaker that need to be identified. During the word identification process, each word from the operating set is compared against the training, pre-recorded set, and the match with the highest accuracy (based on a number of different techniques that are specific to each model's implementation) is selected as the identified word. Pattern matching models have proven successful when the conditions under which the training set is recorded match the operating environment exactly, resulting in limited-vocabulary, speaker-dependent, isolated-word recognition systems. The same pattern-matching algorithms, however, show little tolerance for differences between the operating and training environment conditions. For example, any noise present in the operating environment that differs from the noise in the training environment diminishes the accuracy with which a statistical model identifies the words spoken by the speaker. The accuracy levels can sometimes be retrieved with additional re-training and adaptation. In adverse conditions, however, such as noisy environments or telephone conversations, re-training does not necessarily help performance [6]. Statistical speech perception models also proved more or less inadequate in dealing with speakers having accents or speech disorders and impediments. For these reasons, there is a need

for a different approach in modeling speech perception that will result in systems that are more independent of the environment and speaker. In addition, there has been interest in creating models that would attempt to imitate ways in which humans process speech. The contributions of such a model then, would not only be limited to the industry and production of more robust speech recognition systems, but would stand as a direct quantitative measure of accuracy for phonetic and linguistic theories.

The goal of the Lexical Access From Features (LAFF) project is to develop a speech perception model that would more closely resemble the process by which human listeners are able to extract word sequences from running speech. As a part of the LAFF project, the goal of this thesis is to quantify the acoustic characteristics of nasal consonants and nasalized vowels in American English and incorporate them in an automated speech recognition system. Nasal consonants were chosen because their production and perception has been studied extensively, yet their detection has been a problem for some recognition systems. Nasalized vowels have been included in this study because the production and perception studies for nasal consonants have indicated their presence as important acoustic information in nasal detection.

Once quantified, the acoustic characteristics of nasality would be included in automatic detection systems designed for use in a speaker-independent, continuous-speech environment.

1.2 Acoustic Studies of Speech

The LAFF project is based on the hypothesis that words are represented in memory as sequences of segments, each consisting of bundles of distinctive binary features. These feature value pairs make up a universal set that is used by people

worldwide – any one language uses a subset of features from this set. The feature pairs are such that a change in the value of one feature could result in the production or perception of an entirely different word. For example, while the vocal tract positioning is almost identical in the word pair bat/pat in English, the distinguishing factor between the two initial consonants, and thus words, is whether the vocal folds are vibrating during the consonant production.

Within the feature system, there are two types of distinctive features – articulator-free and articulator-bound [10]. Articulator-free features classify segments into broad classes that can be roughly described as vowels and general classes of consonants by referring to the characteristics of the type of constriction formed in the vocal tract [24]. Articulator-bound features specify which primary articulator is used to make the constriction in the vocal tract and possible secondary articulators that may be involved in the final sound output.

1.2.1 Articulator-free features

By describing the type of constriction formed in the vocal tract, articulator-free features establish one of the broadest classifications between segments – the distinction between vowels and general classes of consonants. Vowels are produced with a relatively open vocal tract and uninterrupted airflow. The acoustic consequence of such a vocal tract configuration is high energy across all formants and continuous movement of the formants throughout the duration of the vowel. The production of a true consonant, in contrast, involves a sequence of movements that produces a narrowing in the vocal tract and that subsequently releases that narrowing, acoustically resulting in two discontinuities in the spectrum – one at the time of closure and another at release.

Because an articulator-free feature is one that has no dedicated articulator to implement it, the narrowing in the vocal tract can be implemented with the lips, tongue blade, or tongue body. Acoustically, we consider articulator-free features as introducing landmarks – the most salient points in an utterance around which information about the underlying distinctive features can be extracted.

1.2.2 Articulator-bound features

Articulator-bound features specify which articulators are active in the vowel or consonant production, and how these articulators are shaped and positioned. There are seven articulators that determine the set of articulator-bound features; they are (1) the lips, (2) the tongue blade, (3) the tongue body, (4) the soft palate or velum, (5) the pharynx, (6) the glottis, and (7) adjustments of the tension of the vocal folds [24]. Each of these articulators can be maneuvered in one or more ways to determine the binary value of the corresponding feature. Because the ways in which the articulators can be manipulated are directly related to the type of constriction made, acoustically we consider articulator-bound features as being reflected in the signal pattern surrounding the landmarks [1]. The landmark times mark the movement and changes in the primary and secondary articulators and for this reason, these times in the signal are most salient in terms of the acoustic characteristics specific to the particular articulator.

1.2.3 Production of Nasal Sounds

A nasal segment is a true consonant in the sense that it is produced by forming a complete closure at some point along the length of the oral region of the vocal tract. The fact that there is a full closure within the vocal tract results in a reduced spectrum

amplitude in the mid- and high-frequency regions, and two acoustic discontinuities – one at the time of the constriction formation (closure) and one at the constriction release (release). The constriction in the oral region is made with the lips, the tongue blade, or the tongue body, similar to other true consonants, such as stops. The principal difference between the nasals and other true consonants is that the velum is lowered and velopharyngeal port open immediately preceding the closure, during the closure, and immediately following the release of the nasal consonant. Because there is an alternative path for the airflow through the nose, there is no pressure increase behind the constriction and the vocal folds will continue to vibrate in a normal manner throughout the closure. This region in the spectrum, during which the vocal folds continue to vibrate despite the full closure and during which the airflow is redirected from the mouth to the nose, is called the nasal murmur. Another characteristic of the nasal segment is that the opening of the velopharyngeal port usually starts during the segment preceding the nasal consonant and the closing can extend into the segment following the nasal consonant. If the preceding or subsequent segment is a vowel, we term the acoustic modifications to the spectrum of the vowel, due to the open velpharyngeal port and additional airflow path through the nasal cavity, as the nasalization of the vowel.

The production of every nasal segment is therefore characterized by some combination of these three events: acoustic discontinuity, nasal murmur, and vowel nasalization. For this reason, estimation of the nasal feature in the speech signal is equivalent to testing for the presence of each of the three events.

In terms of the discussed articulator-free and articulator-bound features, the time of the acoustic discontinuities (one at the time of closure and another at the time of the release) is marked with a landmark, as they denote the time when a narrowing is made or released in the vocal tract. Because the velum is lowered during the production of a nasal segment and the airflow is redirected through the nasal cavity, thus causing no build-up of pressure, the landmark belonging to a nasal segment is that of a sonorant consonant [+consonantal], [+sonorant]. Acoustic properties measured around the landmark time will reflect the lowering of the soft palate which will be captured by the value of the [nasal (soft palate)] articulator-bound feature.

1.2.4 Previous Studies of Nasal Sounds

There has been a large amount of research on both the production and perception of nasal consonants and nasalized vowels. The following are summaries of some of the work done on nasal consonants and nasalized vowels.

- Hattori and Fujimura performed a study in the late 1950's on nasal consonants and nasalized vowels [11]. They reported that the principal features of nasal consonants are a strong resonance located at about 300 Hz, damping of the higher formants, and the presence of an antiformant whose location is dependent on the place of articulation. Despite the difference in the antiformant frequency, however, the overall spectral shape of nasal consonants appeared very similar.

- Using an analog vocal tract synthesizer, House and Stevens found that the major characteristics of vowel nasalization were a weakened and broader first formant,

and overall lower amplitude vowel level compared to non-nasalized vowels [12]. The weaker overall vowel level was a direct consequence of a weaker and broader first formant.

- Hypothesizing that the points of maximal spectral change on either side of the syllabic peak are potential nasal transitions, Mermelstein used four simply extractible acoustic parameters to automatically detect nasals in segmented speech [16]. The four parameters were the relative energy change in the frequency bands 0-1, 1-2, and 2-5 kHz, and the frequency centroid of the 0-500 Hz band. Using multivariate statistics on some 524 transition segments from data of two speakers, Mermelstein achieved 91% correct nasal/non-nasal decision rate. He also noted that the accuracy in a speaker-dependent training system was superior to speaker-independent training.

- In his Master's thesis, Glass reported that the most robust acoustic property of a nasal consonant is a steady, low frequency resonance, which dominates the spectrum [8]. This resonance is characterized by a temporal and spectral stability. By calculating the amount of low frequency energy (below 350 Hz) relative to the energy of the adjacent band (350-1000 Hz), Glass found an effective way of separating nasal consonants from most semivowels. In the same study, Glass stated that the most robust acoustic property of a nasalized vowel was the presence of an extra resonance in the first formant region; depending on the type of the vowel, this resonance might appear above or below the first formant.

- Chen proposed using the parameters A_1-P_1* for non-low and A_1-P_0* difference for other vowel types (in decibels) to capture the spectrum modifications of nasalized vowels [2], [3]. A_1 is the parameter denoting the peak spectrum amplitude of the first formant prominence, P_1 is the spectrum amplitude of a peak near 1 kHz, and P_0 is a spectrum prominence due to a nasal resonance in the range 150-400 Hz. The adjusted acoustic measure of nasalization, A_1-P_1* and A_1-P_0*, were independent of the vowel type.

- In her proposition for a nasal detection module algorithm, Chen combined the parameters for vowel nasalization with a set of parameters that indicate the presence of a nasal murmur [1], [4], [5]. The parameters for the nasal murmur include the location of the lowest resonance, steady-state spectral shape of the nasal murmur as determined by the RMS difference between consecutive frames, and the spectral tilt typical of nasal murmurs as measured by the difference in energies across five frequency bands. The vowel nasalization parameters are those described in [3]. Chen reports that more than 80% of the nasals can be detected correctly with this algorithm.

All of these studies have contributed greatly to the understanding of nasal consonant recognition and perception, but often their results have not been included in speaker-independent, continuous speech recognition systems. Reasons for this include the

lack of a sufficient quantitative form for the data or the fact that many of the collected measurements involved human interpretation and in some cases correction [8].

Using the data from previous studies and building an automated nasal detection module within the LAFF system would provide further insights in the acoustic characteristics of nasal consonants. An automated parameter extraction and decision system would allow for a greater body of naturally spoken data to be examined and quantified. The results of such a system would be directly applicable to the speech recognition systems and would give further insights to phoneticians interested in conducting studies in human speech production and perception.

1.3 Summary

While there is clear evidence that the speech signal is rich in acoustic information regarding its content, there has been little application of speech-specific knowledge to existing speech recognition systems. Consequently, today's systems are mostly speaker-dependent or designed to operate on a very limited vocabulary, which greatly hinders their applicability.

A survey of past work indicates that while acoustic characteristics of speech segments are well studied, they are not always directly applicable to automated speech recognition systems. Their inapplicability is often due to the lack of a sufficiently quantitative form for the presented parameters or because the conducted measurements often require human interpretation and correction.

The primary motivation for this thesis is to examine a number of characteristics and parameters of nasal consonants in American English, and to attempt to incorporate

them in an automatic nasal consonant detection module within the LAFF system. This detection system would operate in a speaker-independent, continuous-speech environment.

The research in this thesis is organized in three stages. The first stage includes automated extraction of parameter values based on previous studies on nasal consonants. As the first step in the automated extraction, Chapter 2 describes the landmark estimation process and performance on a database of utterances. Chapter 3 begins the second stage of processing by introducing the notion of acoustic criteria when processing landmarks for further acoustic characteristics. The following three chapters, Chapters 4, 5, and 6, propose a set of acoustic cues that can separate nasal from non-nasal landmarks, and quantify them in terms of acoustic criteria. In the last stage of nasal module design, Chapter 7 describes the resulting detection system and its performance on the described database. We conclude the thesis with the performance results and future work that would allow this module to operate in a continuous speech environment.

Chapter 2

2.1 Landmark Estimation

In Chapter 1 we discussed the characteristics of a feature-based speech recognition system, and its potential contributions to the industry and future speech studies. We concluded that a feature-based nasal detection module, as illustrated in Figure 2.1, requires three stages of acoustic processing:

1. Estimate the landmark times and types,
2. Use statistically significant acoustic measurements to analyze the signal around the landmark points,
3. Use the values obtained through acoustic measurements to determine whether the secondary articulator involved in the production of the segment is soft palate.

Chapter 2 continues to formulate a feature-based nasal detection module by describing the landmark estimation process on a three-speaker vowel-consonant-vowel (VCV) database of utterances. As the first step in the nasal detection, the landmark estimation output governs the design of the next stage of acoustic processing by specifying which points in the signal should be tested for nasality. Chapter 3 next introduces the concept of acoustic criteria that can be applied to these points to classify them as nasal and non-nasal. In Chapters 4, 5 and 6 we formulate a set of acoustic criteria that can perform this classification as automated algorithms in MATLAB[1]. Defining acoustic criteria for nasality concludes the second processing stage in the nasal detection.

[1] MATrix LABoratory (MATLAB) 6.1.0.450 is property of the MathWorks, Inc.

Lastly, Chapter 7 proposes rules by which these criteria are combined in a nasal detection module, and evaluates their effectiveness on the VCV database.

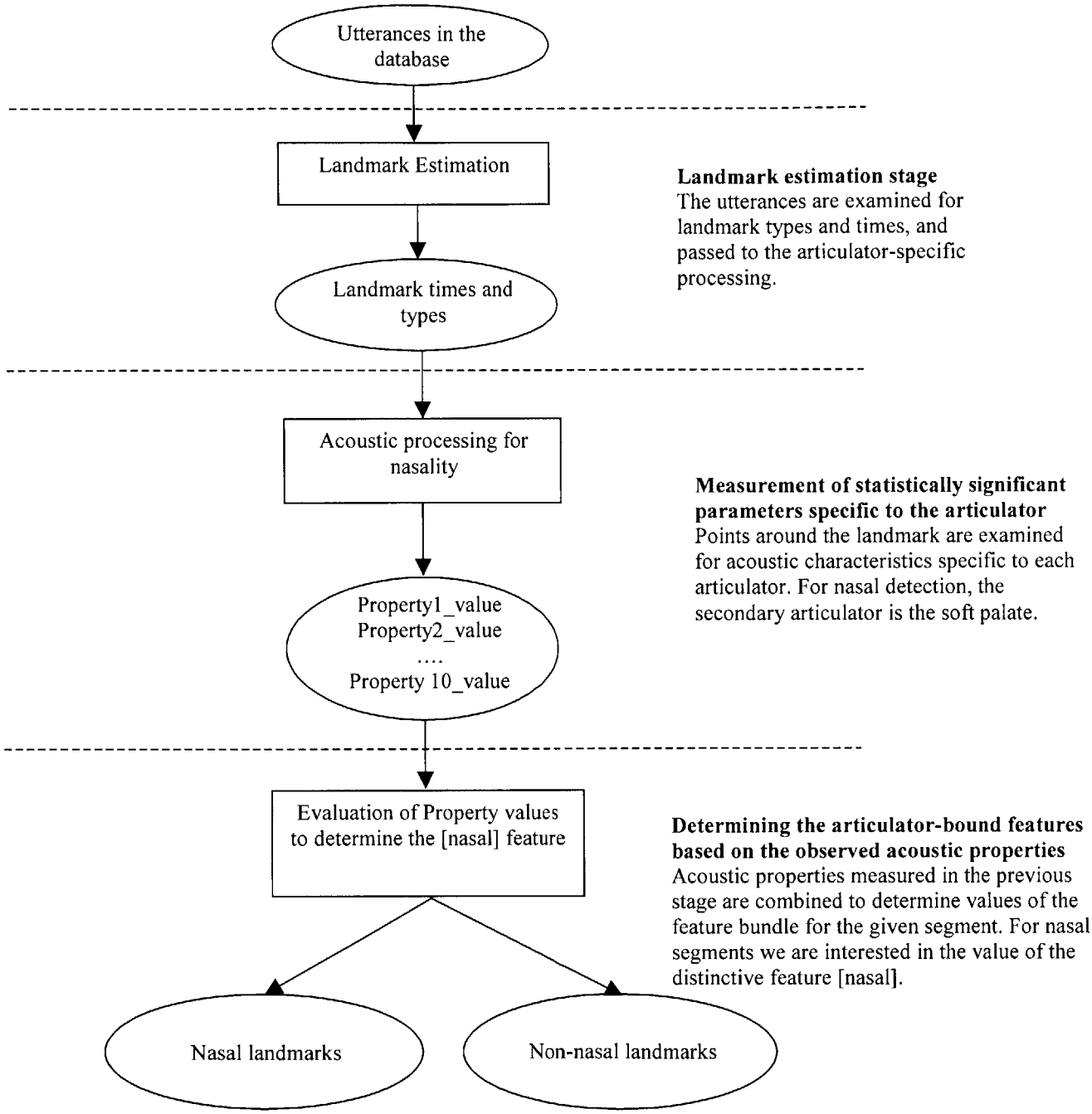


Figure 2.1 – An illustration of the processing required for the nasal detection in a feature-based system. The landmark estimation provides the locus for all articulator-bound feature processing. Once landmarks are estimated, further signal processing is focused on and around the landmarks when determining the primary and secondary articulators. For articulator-bound feature [nasal], the processing examines the acoustic evidence for nasal murmur and/or vowel nasalization by measuring 10 parameters and comparing them against a quantified expectation. Further discussion on the acoustic criteria can be found in Chapters 3 through 6.

Chapter 2 is organized in three sections. The first section describes the VCV database of utterances used in this study. The next section outlines the landmark estimation algorithm and analyzes its performance on the described database. The last section discusses implications of the landmark estimation performance for the design of the nasal detection module.

2.2 Database Description

This study focuses on using a restricted VCV database to study the acoustic characteristics of nasal segments. A restricted database is more likely to have a full set of acoustic cues for nasality as opposed to an impoverished set that might occur in other contexts or in casual speech. Some studies also suggest that stressed syllables are probably articulated with greater care and effort than casual speech, resulting in a more robust acoustic signal and more reliable acoustic features [19].

The analysis database is made with three native speakers of American English (two male and one female) who were between the age of forty and seventy when the recordings were made. The format of each utterance is a vowel at the initial and final position, and a true consonant in the center of the utterance. There are 6 vowels and 26 consonants in the database. Consonants belong to the classes of sonorant (nasal, glide, and liquid) and obstruent (stop, fricative, and affricate) consonants. The database totals 453 utterances, which were passed through an anti-aliasing filter with a cut-off frequency of 7.5 kHz before being digitized at 16 kHz. The 7.5 kHz cut-off frequency allows relevant high-frequency frication noise in female speech to be captured.

The next section outlines the Consonant Landmark Detector (CLD) developed by Liu at MIT's Speech Communication Group, which is used for landmark estimation in this database, and its performance results [13], [14].

2.3 Consonant Landmark Detector

The basis of the LAFF project is the separation of acoustic processing of the speech signal into that of establishing landmark times and types, and using these times as starting points when extracting further acoustic characteristics. The Consonant Landmark Detector (CLD) developed by Liu at MIT is an automated system that provides the landmark processing required for the evaluation of all articulator-bound and some articulator-free features [13]. The CLD module analyzes the digitized speech utterance to produce three types of landmarks as defined by Liu:

1. glottis (g) – time when the vocal folds transition from freely vibrating to not freely vibrating and vice versa,
2. sonorant (s) – time when a sonorant consonant closure is formed or released,
3. burst (b) – stop or affricate bursts and points where aspiration or frication ends due to a stop closure.

The three landmark types determine what further acoustic processing is appropriate for a given landmark. Factors that cause glottal vibration to cease, for example, are buildup of intraoral pressure due to a supraglottal constriction, vocal-fold spreading, or reduction of subglottal pressure. During the nasal segment production, however, the lowering of the soft palate opens an alternative path for the airflow and

there is no change in the intraoral pressure or vocal fold vibration. Because each nasal segment in this database is both preceded and followed by a vowel, utterances with nasal segments will have vocal folds vibrating continually throughout the nasal production. For these reasons, [g] landmarks in a VCV database do not appear to mark an acoustic change in the signal that could result from the soft palate movement. Similarly, the signal around [b] landmarks is expected to show a silence interval followed by an abrupt increase in energy at high frequencies. This property of [b] landmarks suggests that nasal detection should not focus around these landmark points. Unlike the previous two landmark types, however, a sonorant landmark is an acoustic manifestation of a sonorant closure or release in Liu's detection scheme. Consequently, a portion of the signal around this type of landmark is a prime candidate for examining the signal for soft palate activation and nasal production. Figure 2.2, adopted from Liu, illustrates the landmark tree implemented by the CLD. The diagram also shows the dependence between the articulator-bound feature processing and landmark type.

The subsequent sections in this chapter focus mainly on the calculation and performance of sonorant landmarks because of their relevance to the nasal detection. The CLD used in landmark estimation and described below is unchanged from Liu's doctoral thesis [13]. The reader can refer to the thesis for all details not included in the overview.

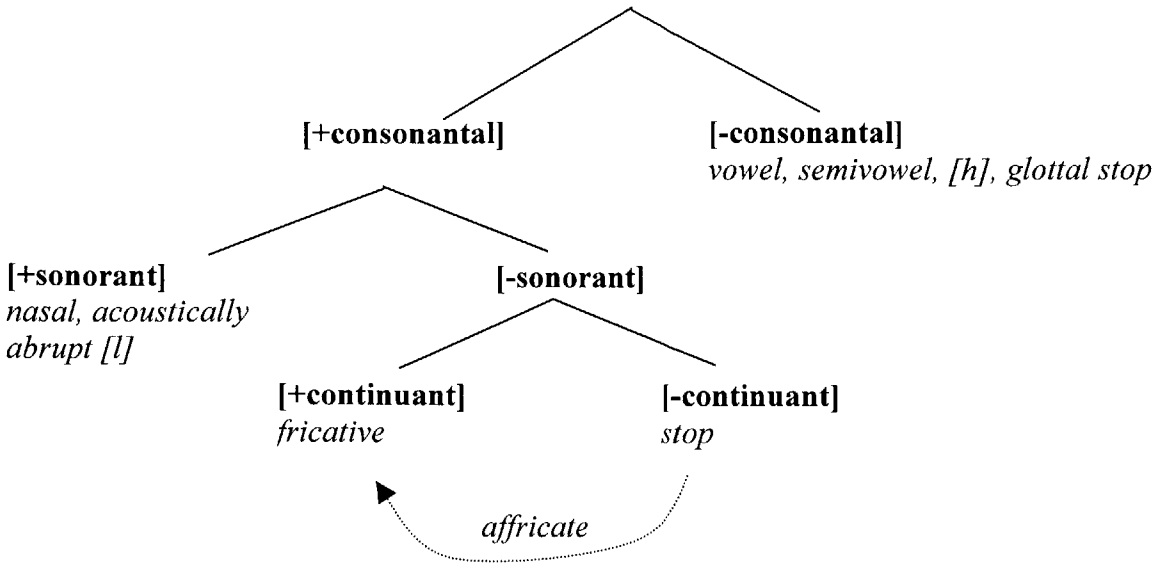


Figure 2.2 – The landmark tree adopted from Liu. From this diagram, it is evident that the articulator-bound feature processing is highly dependent on the information from the CLD in selecting the appropriate acoustic analysis. For nasal detection the landmarks of interest are [+consonantal], [+sonorant].

2.3.1 Overview of the sonorant landmark calculation

Calculation of sonorant landmarks can be separated in general processing common to all landmark types and specific tests for sonorant cues. General processing includes computation of the broadband spectrogram using the short-time processing techniques, with a 6 ms Hanning window taken every 1 ms and a 512-point DFT. The resulting spectrogram is divided into the following six frequency bands before being passed through fine and coarse processing:

Band 1: 0.0 – 0.4 kHz
Band 2: 0.8 – 1.5 kHz
Band 3: 1.2 – 2.0 kHz
Band 4: 2.0 – 3.5 kHz
Band 5: 3.5 – 5.0 kHz
Band 6: 5.0 – 8.0 kHz.

Frequency bands are chosen such that Band 1 monitors the presence or absence of glottal vibration. Bands 2-5 are used to detect spectral changes of sonorant consonants and

onsets/offsets of aspiration or frication noise associated with stops, fricatives, or affricates. Band 6 spans the remaining frequency range and is used in combination with other bands to detect silence intervals in stop consonants. Following the spectrogram calculation, energy changes in the six bands are found using a two-pass strategy. Both passes employ the same processing steps except that the first pass uses coarse parameter values to find the general vicinity of a spectral change and the second pass uses fine parameter values to localize it in time [13]. Once the 6-band energy is computed, a 6-band rate-of-rise (ROR) is found by taking an overlapping dB first difference of the energy in each band. Using a peak-picking algorithm originally described by Mermelstein [16], Liu finds ROR ±peaks whose absolute value is greater than 9dB for coarse processing, and 6dB and 9dB for fine processing. For details of the algorithm, the reader should refer to Liu's doctoral thesis. Detection of energy peaks as a function of time concludes the general processing stage common to all landmark types.

To find sonorant [s] landmarks, the CLD only considers voiced regions of the utterance. Because [g] landmarks are indicators of the time when glottal vibration turns on or off as determined by the energy change in Band 1, sonorant landmarks can only exist in the regions of the signal bounded by a +g landmark on the left and –g landmark on the right[2]. Within the voiced region, peaks in Bands 2-5 are grouped based on the sign and temporal proximity; that is all peaks or dips that happen somewhat coincidentally make up a group that is passed to the next processing stage. There are usually a number of such groups within each utterance. In each group, the largest peak or dip is termed a ***pivot*** and considered a likely candidate for a sonorant landmark. The information

[2] For further information regarding the [g] landmark estimation and analysis, the reader can refer to Liu's doctoral thesis.

regarding the remaining peaks in the group is used when determining whether the pivot passes the landmark requirements. For ease of understanding, Figure 2.3 presents the visual output of the CLD for utterance [ɑnɑ]. The broadband spectrogram tops the figure, followed by the plots of energy in each frequency band as a function of time. The band number is located to the left of each plot. The added notation in Figure 2.3 highlights two groups of energy fluctuations that are considered to occur somewhat coincidentally: the first group has 3 energy peaks and the second 4 dips in Bands 2-5. The largest peak in each group across the 4 frequency bands is a pivot, also noted on the diagram.

To become a sonorant landmark, each pivot needs to pass three criteria of sonorant regions labeled steady-state, abruptness, and staggered peaks criteria. Table 2.1 summarizes the theoretical basis and expectation for each test.

Name of test	Theoretical basis
Steady-state	After the primary articulator has made a complete closure, the vocal folds continue to vibrate and the vocal tract shape is relatively constant, resulting in a relatively unchanged low frequency content during the constricted interval.
Abruptness test	Acoustic manifestation of a sonorant release or closure is a rapid change in the F2 to F4 range – decrease in the energy for sonorant closure and increase for a release.
Staggered-peak test	Another measure of high-frequency abruptness for sonorant consonant states that peaks in each group must occur somewhat coincidentally with each other and their pivot.

Table 2.1 – Overview of the criteria used in the sonorant landmark estimation. Summary of the requirements that each pivot has to pass before it is promoted to a sonorant landmark. The table illustrates the theoretical basis for each of the tests in the CLD.

Pivots that pass all three criteria become sonorant landmarks. The three criteria also determine whether the landmark is located at the time of a sonorant closure or release, while the sonorant landmark time is the same as the time of the pivot. Pivots that pass the abruptness and staggered-peak criteria, but not the steady-state, are classified as the specially introduced fourth landmark type – vocalic [v] landmark.

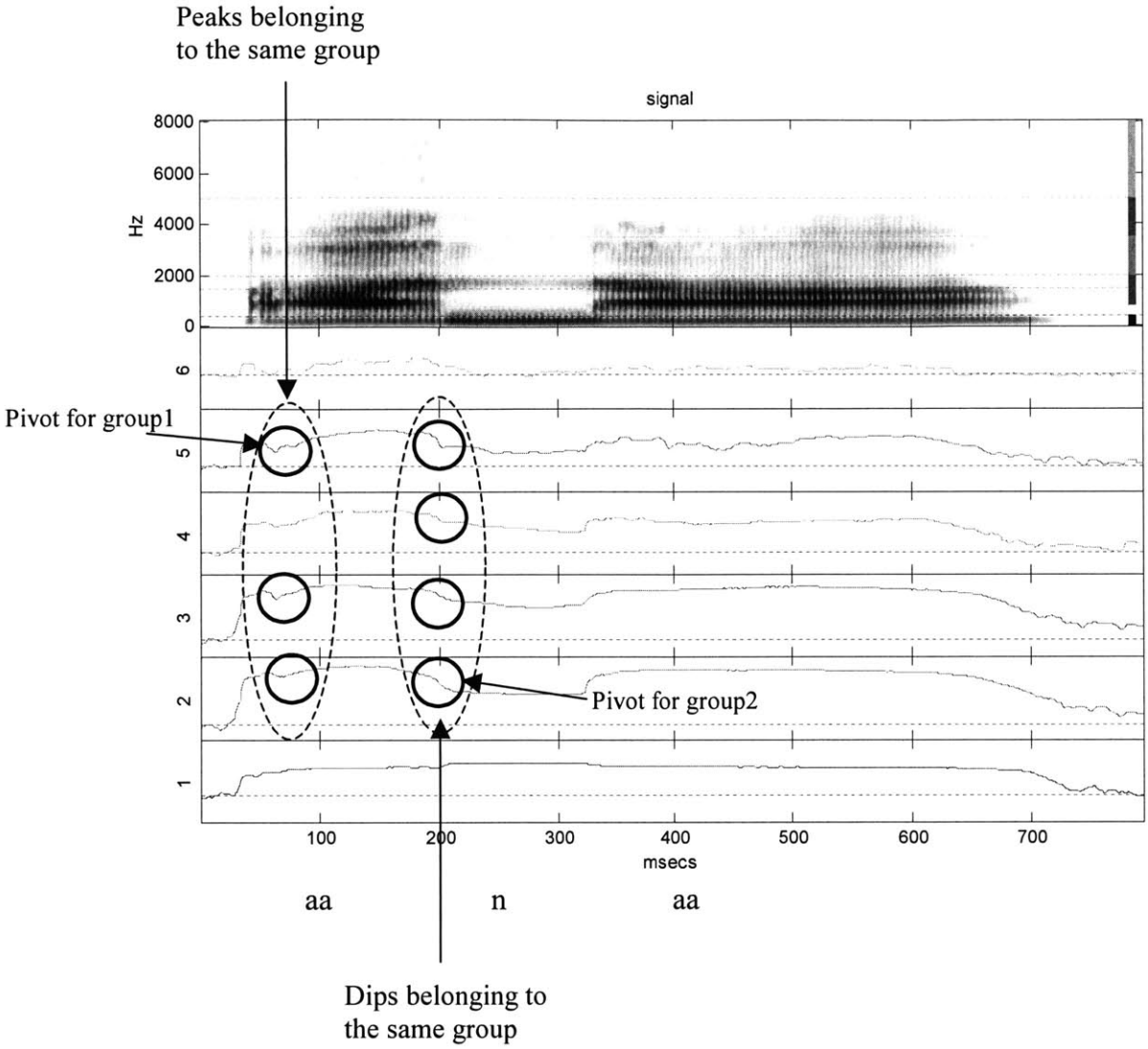


Figure 2.3 – The visual output of the CLD gives the broadband spectrogram of the utterance and the energy in each frequency band plotted as a function of time. The frequency band number is noted on the left of each plot – information regarding the band ranges can be found in the text. After the energy of each frequency band is analyzed in terms of somewhat rapid energy changes, the peaks and dips are grouped based on their temporal proximity and sign. The largest peak/dip in each group is named a pivot and considered a likely candidate for a sonorant landmark.

The origin of all [v] landmarks in utterances processed by the CLD is thus a pivot that passed two of three requirements for a sonorant landmark. The remaining pivots that do not meet the sonorant or vocalic landmark criteria are discarded. Figure 2.4 summarizes the process by which pivots are analyzed for sonorant landmark

characteristics. Once all pivots have been examined, the landmark types and times are saved to be written to a text file. This point concludes the sonorant detection within the CLD.

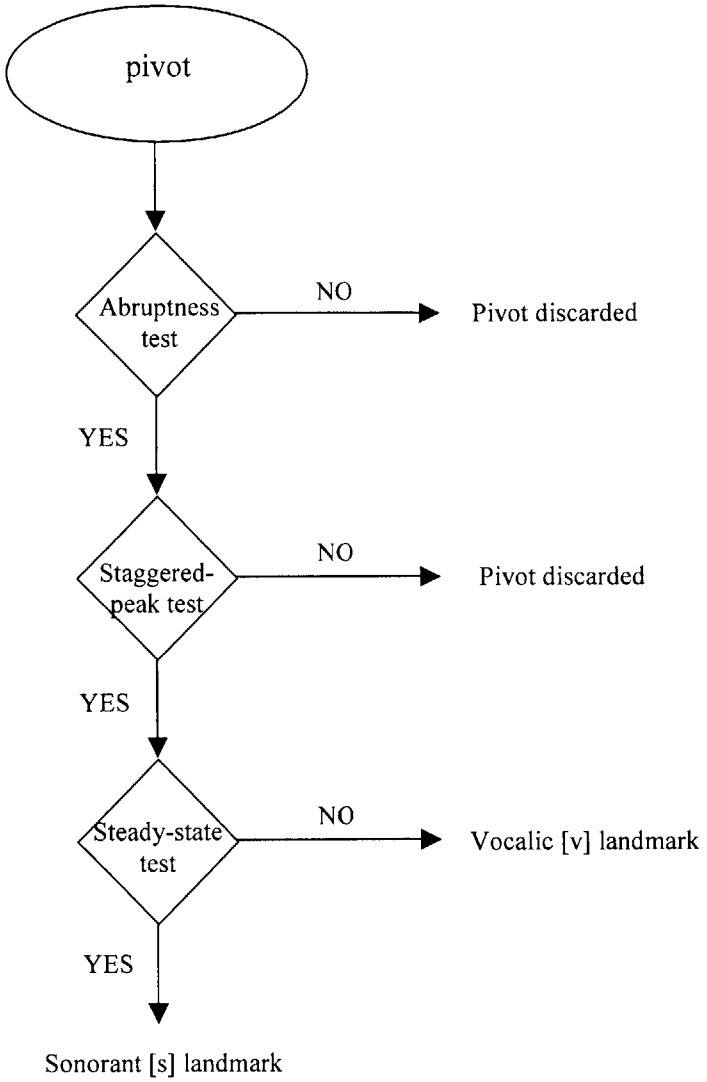


Figure 2.4 – Illustration of the decision process used to determine whether a pivot is promoted to a sonorant landmark. Table 2.1 summarizes the theoretical basis for each test.

2.3.2 Sonorant landmark performance on the VCV database

Liu chooses a sonorant landmark to mark the time in the utterance when a sonorant closure or release is made [13]. The sonorant class of consonants in Liu's

detection scheme includes only nasals and acoustically abrupt [l] segments; glides and liquids are classified as semivowels and described as [-consonantal]. Assuming that [m], [n], [ŋ], and [l] are all acoustically abrupt in a VCV database, each of the above segments should have two landmarks: [-s] to designate the time of the sonorant closure and [+s] of sonorant release. Although sonorant landmarks in a VCV database are expected to appear in pairs, the CLD does not enforce this property. A sonorant closure or release can and often does exist without a corresponding sonorant pair in a VCV database due to CLD performance errors or in instances when closures and releases are not acoustically abrupt. For this reason, we adopt the sonorant landmark (closure or release) as the basic unit when evaluating the accuracy of the sonorant landmark detection in this database.

The described VCV database of 3 speakers and 6 vowels has 72 sonorant segments (definition for a sonorant consonant is adopted from Liu and includes nasals and acoustically abrupt [l] segments). With the utterance [iŋi] missing in the database for one speaker due to a corrupted file, this number is adjusted to 71 acoustically abrupt sonorants – 53 nasals and 18 acoustically abrupt [l] segments – each with an expected landmark at the sonorant closure, [-s], and release, [+s]. We thus expect 71 sonorant closures and releases for a total of 142 sonorant landmarks. Table 2.2 contrasts the actual against the expected performance.

Further analysis into the output of the CLD reveals two types of error:

1. Some sonorant closures and releases are undetected by the CLD – we will refer to these as ***deletions*** in the rest of the study,
2. Some sonorant landmarks are placed at non-sonorant segments, which we label ***insertions*** for the remainder of the study.

		-s	+s	Total
Actual number	Nasal	31	39	70
	Abrupt [l]	3	11	14
	Other	121	46	167
Expected number	Nasal	53	53	106
	Abrupt [l]	18	18	36
	Other	0	0	0

Table 2.2 – An overview of the CLD performance results for the described VCV database. This table indicates that some of the sonorant closures and releases were undetected by the CLD, while some sonorant landmarks were erroneously placed within non-sonorant segments.

In order to fully analyze each of the error types and rates, we require a means to further classify estimated sonorant landmarks based on the location within the utterance where they occurred. Hand-classifying each sonorant landmark as

1. true positive – placed at a closure or release for a nasal or [l],
2. or false positive – placed within a vowel, semivowel, at a vowel-semivowel/semivowel-vowel (VS/SV) boundary, or vowel-obstruent/obstruent-vowel (VO/OV) boundary,

produces the desired detection rates. The sub-classification of true and false positives also gives insight in the structure of detections and insertions. Table 2.3 shows the number of true and false positive sonorant landmarks distributed across the six groups. The list of utterances, landmark times and types, and their classification into one of the six groups can be found in Appendix A. If we define the detection rate to be

$$DetectionRate = \frac{Detections}{Expected\_Sonorant\_Landmarks} \times 100\%, \quad (2.1)$$

the detection rate of the sonorant landmarks for this VCV database is 59.2%.

	Comment	-s	+s	Total [s] landmarks	Expected [s] landmarks
Nasal	True positive	31	39	70	106
Abrupt l		3	11	14	36
Within Vowel	False positive	77	21	98	--
VS/SV boundary		5	13	18	--
VO/OV boundary		39	11	50	--
In semivowel		0	1	1	--
Total	**True positive**	**34**	**50**	**84**	**142**
	False positive	**121**	**46**	**167**	--

Table 2.3 – This table gives insight into the structure of the false positives and detection rates. Each sonorant landmark produced by the CLD can be hand-classified as a true or false positive. True positives (detections) are landmarks at closures and releases for nasals and [l]s. While each sonorant segment in this database is expected to have exactly two sonorant landmarks – one for the closure and one for the release - the actual results show that sonorant landmarks had anywhere between zero and two landmarks. False positives can be exhaustively classified in four categories, depending on whether they were inserted in a vowel or semivowel segment, at a vowel-semivowel/semivowel-vowel boundary, or vowel-obstruent/obstruent-vowel boundary.

Qualitatively such a low detection rate means that for each detected sonorant landmark, almost one other is missed. Detection rate is higher for nasals than for acoustically abrupt [l] segments, which is in agreement with Liu's observation for the original CLD databases [13]. Furthermore, based on a limited number of tokens, it appears that sonorant closures are better detected when they are adjacent to certain vowel types. To illustrate this, Table 2.4 shows the number of detected sonorant landmarks based on the adjacent vowel. Although more speakers/tokens are needed in order to make a conclusive claim, it appears that the sonorant landmark detector does not perform well if the sonorant closures are adjacent to a nonlow back vowel.

	ɑ	a	ɛ	i	o	u
-s	8	6	9	9	2	0
+s	9	10	9	6	10	6
Total	17	16	18	15	12	6
Detection Rate (%)	**17/24 (70.8)**	**16/24 (66.7)**	**18/24 (75.0)**	**15/22* (68.2)**	**12/24 (50.0)**	**6/24 (25.0)**

*The number of expected sonorant landmarks for vowel [i] is different because of the missing utterance [iŋi] for one speaker.

Table 2.4 – Distribution of detected sonorant landmarks based on the adjacent vowel. For example, a detected [-s] landmark at a sonorant closure before the vowel [ɑ] would contribute one token to the [–s] entry under the vowel [ɑ] in the table. This sub-classification illustrates the disparity between the detection rates of sonorant landmarks adjacent to non-low, back vowels compared to the rest.

A plausible explanation for the poor performance is that the low frequencies of the first and second formants, which are characteristic of nonlow, back vowels, cause them to have relatively weak energy at higher formants. The low energy at frequencies above the second formant in the vowel segment in turn produces a rather smooth transition from the vowel to the sonorant closure that does not satisfy the required energy abruptness criterion. Failure to meet this criterion precludes the pivot from further landmark consideration. In support of this hypothesis is the trend of the detection rate – the lowest detection rate is for the vowel [u], which has the lowest frequencies for F_1 and F_2. If this hypothesis is true, the sonorant landmark detector in the CLD could use the information regarding the adjacent vowel type to calibrate its criteria. This calibration could be implemented as a second pass in the landmark estimation: once the vowel types and times are determined, the sonorant landmark estimation within the CLD could lower the energy threshold requirement for pivots adjacent to the nonlow, back vowels and reexamine them.

Examination of insertions, and in particular the disproportionate number of sonorant landmarks inserted in vowel segments, reveals that about 50% of vowel

insertions are either at the beginning of the utterance, for the onset of voicing, or at the end of the utterance as the voicing ceases and the speaker's voice becomes aspirated. For continuous and spontaneous speech with longer utterances, we expect the vowel insertions as a percentage of total insertions to decrease significantly. Insertions at obstruent segments, in semivowels, and those placed near the center of a vowel do not appear to be specific to the VCV database used in this study.

2.4 Designing the system in terms of the [s] landmark performance

The analysis so far described the detection rate and error types of the sonorant landmark detection and CLD in isolation from the nasal detection module. This section characterizes the implications they have on the design of the nasal detection module. Design concerns are organized in two sections based on the error type.

2.4.1 Design considerations of the nasal detection module due to deletions

Because a feature-based speech recognition system relies on estimated landmarks for further articulator-bound feature processing, low sonorant landmark detection rate of the CLD presents a substantial problem for the nasal detection. With a deleted landmark, feature-based nasal detection has no information that an acoustically abrupt change occurred and will not examine the signal for further acoustic characteristics. A methodical approach to this problem suggests two possible solutions:

1. Modifications to the CLD algorithms are able to raise the detection rate to a sufficiently high level such that the nasal detection module need not address the case of deleted landmarks. One such enhancement, as we have already

speculated, could include lowering the energy abruptness threshold or restricting the threshold to certain bands for a pivot based on the type of the adjacent vowel.

2. The design of the nasal detection module accounts for deletions by evaluating possible nasal contexts in the signal other than the landmark points. The additional processing may include testing [v] in addition to [s] landmarks because of their origin as sonorant landmark candidates that failed the steady-state criterion.

In this study we choose to address the problem of low sonorant detection rate from within the nasal module. The reasoning behind this approach is that several articulator-bound modules within the LAFF project are either already developed or in progress based on the existing CLD developed by Liu. Customizing the CLD for sonorant landmarks would thus require that the nasal detection module include its own version of the CLD when added to the LAFF system. Following in the same fashion, if more articulator-bound feature modules were to include their own customized CLD, the separation between the articulator-free and articulator-bound feature processing that is the basis of the LAFF project would no longer exist. Sonorant landmark estimation in addition depends on [g] landmarks to establish voiced and unvoiced regions in the utterance. Modifications could thus require changes in the processing of both [g] and [s] landmarks to enhance its performance.

Working with the existing CLD algorithm, we change the focus of our analysis from sonorant landmarks to pivots and examine whether pivots can be used to compensate for the low detection rate of sonorant landmarks. As described earlier, pivots

have to pass a set of three criteria to become sonorant landmarks. Based on which criteria they satisfy, pivots are either promoted to sonorant [s] or vocalic [v] landmarks, or discarded. This decision process raises the following questions as suggested by Chen [1]: How many deletions originate as a pivot that was erroneously promoted to a vocalic landmark or discarded based on the three criteria, and how can we use this information to compensate for the low sonorant detection rate?

The first step in answering these questions is to hand-label each sonorant closure and release in the database, and compare these times against the pivots produced by the CLD. Table 2.5 outlines the number of sonorant closures and releases that were examined by the CLD as pivots and their final classification.

Sonorant type	Pivots promoted to [s] landmarks	Pivots promoted to [v] landmarks	Discarded Pivots	Total pivots	Missing pivots or pivots replaced by [g] landmarks	Expected pivots
Nasal	70	7	26	103	3	106
Abrupt [l]s	14	2	18	34	2	36
Total	**84**	**9**	**44**	**137**	**5**	**142**

Table 2.5 – The pivot analysis suggests that pivot detection rate for sonorant closures or releases is significantly higher than that of sonorant landmarks. If pivots are used as possible nasal context in an utterance instead of sonorant landmarks alone, 137 of existing 142 sonorant closures and releases would be examined for possible nasal characteristics.

Table 2.5 shows that 97.2% of nasal closures and releases, and 94.4% of abrupt [l]s, are examined as pivots during the CLD processing. For nasals, approximately 7% of examined pivots are falsely promoted to vocalic [v] landmarks and 25.2% are discarded. Even larger percentage of pivots is erroneously rejected for abrupt [l]s; about 6% of

abrupt [l]s are classified as [v] landmarks, while about 53% of pivots are falsely discarded. For the remaining <4% of nasals and laterals that did not have a pivot, the majority were a consequence of a CLD processing rule which requires that no pivots exist in the proximity of [g] landmarks. In these utterances the placement of a [g] landmark close to the sonorant closure or release caused that portion of the signal not to be included in the pivot analysis, ultimately resulting in a deleted sonorant landmark. In one instance a missing pivot was caused by an acoustically non-abrupt transition between the vowel [u] and a nasal. Pivot analysis data can be found in Appendix B.

Selecting a pivot instead of a sonorant landmark as a possible nasal context in the nasal detection module guarantees that >96% of all sonorant closures and releases will be examined for nasality. The drawback to this approach is that in the VCV database of 453 utterances and 71 sonorant segments, the CLD examines 1,483 pivots, thus significantly increasing the amount of processing needed for the nasal detection. In order to avoid this increase in the computational requirement, we require that the acoustic criteria for nasalization be ranked based on their effectiveness in separating nasal from non-nasal pivots. By employing the most effective criteria first, we minimize the required computational power by rejecting most non-nasal pivots early without calculating parameters used for the remaining criteria. More on the implementation of the ranked processing scheme can be found in Chapter 7.

2.4.2 Design considerations of the nasal detection module due to insertions

Figure 2.2, adopted from Liu, suggests that following the landmark estimation sonorant landmarks need only be separated into those belonging to nasals and

acoustically abrupt [l]s. The high insertion rate of sonorant landmarks within non-sonorant segments requires that nasal detection also accounts for and rejects false positives. Including the classification of the false positives from previous sections, Figure 2.5 alters the original landmark tree implemented by Liu to account for the observed performance characteristics of the sonorant landmark detection.

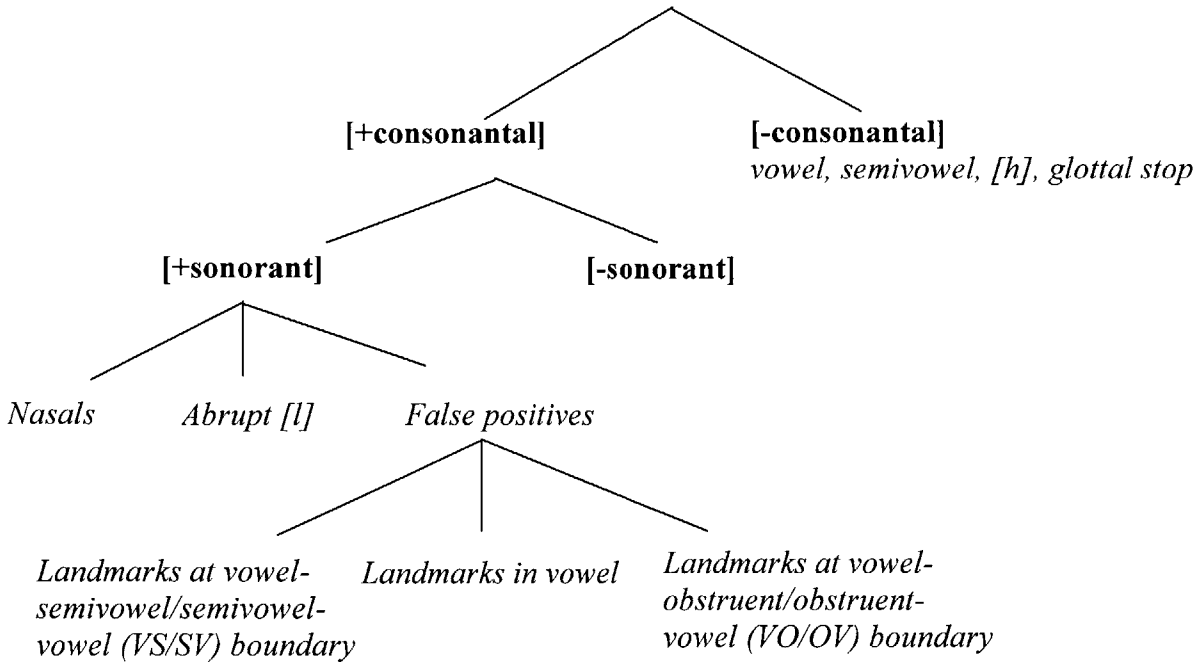


Figure 2.5 – The landmark tree proposed by Liu is modified to account for the sonorant landmark insertions seen in the CLD performance. Insertions of sonorant landmarks within non-sonorant segments require that nasals are detected from acoustically abrupt [l]s and false positives, which include sonorant landmarks placed in vowel, semivowel, and obstruent segments. The variety of non-nasal groups suggests that a larger number of parameters need to be used to successfully separate nasal from non-nasal sonorant landmarks.

The inserted landmarks effectively expand the number of classes under the sonorant landmark group to nasals, abrupt [l]s, vowels, semivowels, and obstruents. With such a wide variety of segments and boundaries, the design of the nasal detection module requires a number of measurements to successfully separate nasals from the remaining

groups. In Chapters 3-6, we propose and test the effectiveness of 10 acoustic cues in separating nasal from non-nasal sonorant landmarks. Because these parameters are based on a number of past studies, the focus of our discussion is on their formulation within a set of automated algorithms in MATLAB that are used to extract them. The question that we attempt to answer is whether we can select a subset of cues that can be formulated as automated algorithms that will successfully separate nasal from non-nasal landmarks. Selecting a subset of these measurements in Chapters 3 through 6, and combining them into a feature-based module in Chapter 7 concludes the goal set out at the beginning of this study, which is to design a successful automated feature-based nasal detection module.

2.5 Summary

Chapter 2 describes the landmark estimation on a vowel-consonant-vowel (VCV) database as the first processing stage in the nasal detection module. In landmark estimation we use Liu's Consonant Landmark Detector to produce three types of landmarks: [**g**lottis], [**s**onorant], and [**b**urst] landmarks. Based on the acoustic event described by each landmark type, we focus further discussion on sonorant landmarks as prime candidates for nasality.

The performance of the CLD for this particular database is characterized by a low detection rate of sonorant closures and releases, and a high rate of sonorant landmarks inserted within non-sonorant segments. Sonorant segments are detected at a rate of 59.2%, where nasals are better detected than abrupt [l]s. Both findings are in accordance with the results reported in Liu. Inserted landmarks can be exhaustively classified in four

groups; sonorant landmarks inserted within vowel and semivowel segments, at vowel-semivowel/semivowel-vowel, and vowel-obstruent/obstruent-vowel boundaries.

The low detection rate of sonorant landmarks prompts the discussion of pivots as potential candidates for nasality. This analysis concludes Chapter 2 with two recommendations:

1. The final nasal detection module should focus its processing on ***all*** pivots examined by the CLD – combined they have been shown to capture >96% of all sonorant closures and releases in our VCV database,
2. The acoustic criteria should be designed in such a way as to minimize the computational power required by the nasal detection module.

Chapter 3

3.1 Acoustic Criteria as the Basis for Nasal Detection

In Chapter 2 we described the first stage in the three-stage model of a nasal detection system by detailing the process of landmark estimation on a vowel-consonant-vowel (VCV) database of utterances. In the following four chapters, we use the 250[1] nasal and non-nasal sonorant landmarks estimated and hand-labeled in Chapter 2 to design the second processing stage – formulating a set of acoustic criteria that can successfully classify these landmarks as nasal or non-nasal through automated algorithms. In other words, in the remaining chapters we focus on finding promising acoustic cues that can successfully classify each sonorant landmarks as nasal or non-nasal. The design process is divided between four chapters for clarity. In Chapter 3 we describe the meaning and significance of the acoustic criteria in nasal detection and general processing of the signal and data analysis common to all chapters. Chapters 4, 5, and 6 propose a set of acoustic cues and examine their effectiveness at detecting three events that are associated with the production of nasal segments in American English – vowel-nasal boundary, nasal murmur, and nasalized vowel. Each chapter concludes with a definition of an acoustic criterion or criteria for the specific event and an analysis of its performance when applied to the set of estimated sonorant landmarks from Chapter 2. Chapter 7 combines the results of Chapters 2 through 6 by applying the formulated acoustic criteria on the 1483 pivots estimated for this VCV database and examining their performance in separating nasal from non-nasal pivots.

---

[1]The semivowel segment group of estimated sonorant landmarks from Table 2.3 is not used in the measurements presented in the next three chapters – analysis on a single-member group is thought to have limited significance for the overall results.

3.2 What are acoustic criteria?

When testing a pivot or landmark for nasality, the nasal detection module analyzes the portion of the signal immediately preceding and following the pivot or landmark time. For a pivot located at the time of a nasal closure in a VCV database, for example, the preceding signal portion is a vowel, and is specifically expected to exhibit signs of vowel nasalization. The portion of the signal following the pivot time, on the other hand, follows the nasal closures and is expected to show characteristics of nasal murmur described in Chapter 1. To classify a pivot as ***nasal*** then, is to confirm that the signal surrounding the pivot time exhibits some combination of the nasal boundary, nasal murmur, and vowel nasalization cues. A ***non-nasal*** pivot, by the same token, fails to exhibit such acoustic characteristics. A non-nasal pivot can be a true sonorant landmark or pivot placed at the closure of an abrupt [l], or a false positive – a sonorant landmark or pivot located within a vowel or semivowel segment, at a vowel-semivowel/semivowel-vowel (SV/VS) or vowel-obstruent/obstruent-vowel (VO/OV) boundary. The regions surrounding these pivots will have diverse acoustic characteristics, which will generally differ from those of the vowel nasalization and nasal murmur. Figure 3.1 illustrates the utterance [ama] with two pivots – the first located within the initial vowel at 76ms and the second pivot located at the time of the nasal closure, around 184ms. From the diagram, it is evident that characteristics of the signal on at least one side of the pivot time differ for a nasal and non-nasal pivot. The goal of the acoustic criteria is to capture this acoustic difference.

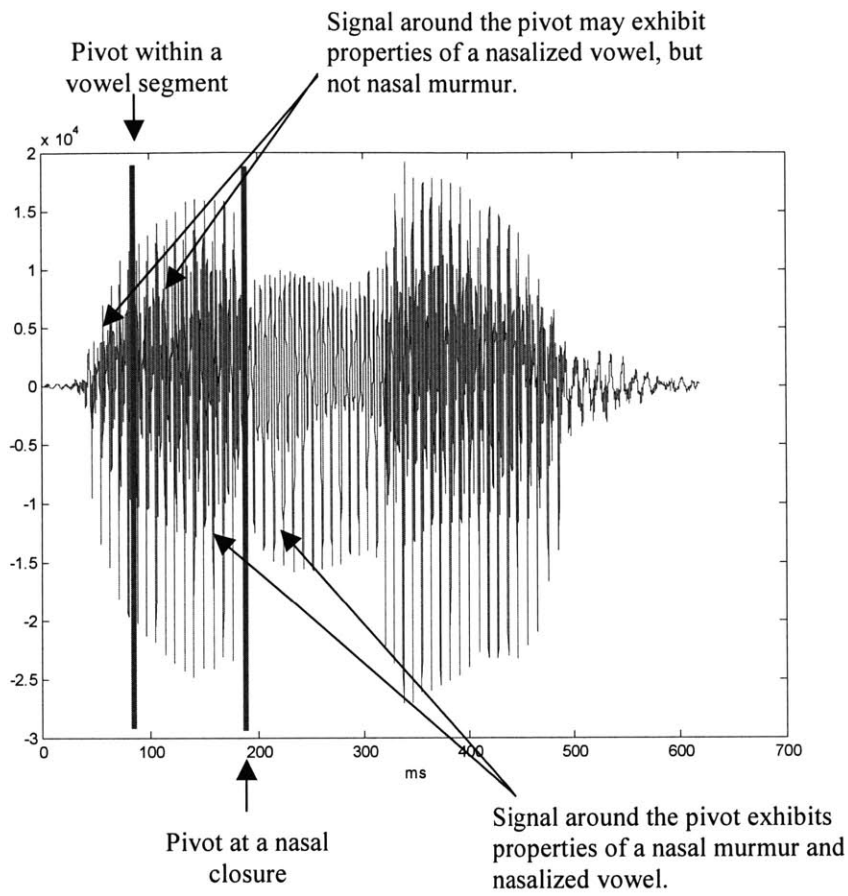


Figure 3.1 – This figure contrasts a non-nasal and nasal pivot. The non-nasal pivot is located within a vowel segment, around 76 ms, and while the signal around it may exhibit properties of a nasalized vowel, it fails to exhibit characteristics of a nasal murmur. The nasal pivot at 184 ms, on the other hand, exhibits characteristics of both – the nasalized vowel precedes the pivot time and the nasal murmur characteristics are reflected in the portion of the signal following the pivot time.

3.3 Acoustic criteria around the landmark point

Sonorant landmark determines what portion of the signal around the landmark point should adhere to the specific acoustic criterion. The sign associated with the sonorant landmark indicates the general trend in the change of the spectral energy in Bands 2-5 – a [-] sonorant landmark indicates decreasing energy due to a more constricted vocal tract, while a [+] landmark implies increasing energy most often associated with a constriction release. Each sonorant landmark thus divides the signal

into a region with higher and lower energy levels. To detect nasals in a set of nasal and non-nasal sonorant landmarks[2] estimated on a VCV database is to verify that the signal around the landmark point exhibits a combination of the following properties:

- The change in energy displays characteristics of a vowel-nasal or nasal-vowel boundary,
- The decrease in energy is due to a nasal closure – consequently, the region of the signal with a lower energy level shows properties of the nasal murmur,
- The portion of the signal adjacent to the landmark with a higher energy level displays properties of a nasalized vowel.

The role of the acoustic criteria for nasality is to test the signal around the sonorant landmark and determine whether it shows characteristics of the above three events. As such, acoustic criteria can be divided into a criterion for the nasal boundary, nasal murmur, and nasalized vowel depending on whether they examine the acoustically abrupt transition, or region of lower or higher energy level.

An acoustic criterion for nasality usually consists of an acoustic cue or a set of acoustic cues, and the quantified expectation regarding their behavior. The cues selected for an acoustic criterion show some special property when measured in the portion of the signal around a nasal that differs from making the same measurements around a non-nasal segment. Figure 3.2 illustrates an example of an acoustic cue, AC1, for the nasal murmur. When measured in the portion of the signal following a nasal [-s] landmark, the

[2] Because our discussion in Chapter 4,5, and 6 is based on a set of 250 nasal and non-nasal sonorant landmarks from Chapter 2, we reserve the discussion on how pivots convey the expectation of acoustic characteristics for Chapter 7.

range of values for this cue is [0,0.5]. Measured in the signal following a non-nasal [-s] landmark, however, the same cue exhibits a significantly different behavior, with the values falling in the [0,3] range. Knowing that AC1 will behave differently when measured at nasal and non-nasal sonorant landmarks allows us to incorporate the knowledge of its quantified behavior within the automated nasal detection module. If measuring AC1 in the portion of the signal following a [-s] or preceding a [+s] landmark produces a value of 2, the nasal detection module can classify that portion of the signal as not belonging to a nasal murmur. If the produced value is less than 0.5, we require additional acoustic cues to determine whether the landmark is indeed a nasal. The expectation that the value of AC1 does not exceed 0.5 or 1 for a nasal landmark is an example of an acoustic criterion. Interpretation of box plots, such as the one in Figure 3.2, can be found in Section 3.4.1.

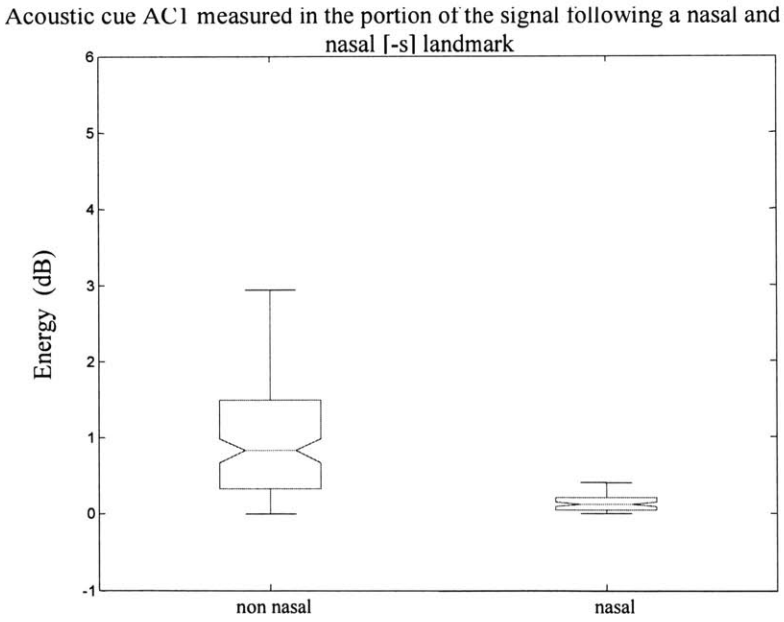


Figure 3.2 – An example of a significant cue AC1 measured in the portion of the signal following nasal and non-nasal [-s] landmarks. The consistently differing values of AC1 when measured around nasal and non-nasal sonorant landmark suggest that its quantified behavior can be used as an acoustic criterion in deciding whether sonorant landmarks show the nasal murmur characteristics.

3.4 Our approach to the acoustic criteria analysis

Acoustic cues characteristic of the nasal boundary, nasal murmur, and vowel nasalization have been studied to a varying degree in the past. On the one hand, Chen fully formulates an acoustic criterion for the nasal murmur that included all estimated nasal sonorant landmarks and rejected 54.5% of non-nasal landmarks for nasal murmur in the original database of utterances [1]. The acoustic cues for the nasal boundary, on the other hand, do not appear to have a sufficiently quantified form that can function as the basis for an acoustic criterion in an automated module. In addition, all criteria and cues from past studies quote values obtained through hand-measurements that often include human interpretation and correction. In order to incorporate these results in the automated nasal detection module, in the next three chapters we:

1. Formulate all available criteria and cues in terms of automated algorithms in MATLAB,
2. Analyze them on the 250 hand-labeled nasal and non-nasal sonorant landmarks by:
    - Examining differences in the distribution of each cue when measured around nasal and non-nasal sonorant landmarks,
    - Analyzing performance of the existing acoustic criteria and suggesting possible improvements.

The values of extracted cues are thus analyzed in terms of the past studies and through statistical analysis of the difference in the extracted values for the nasal and non-nasal group of sonorant landmarks. The next section briefly explains the methods used for the statistical data analysis.

3.4.1 Analysis of variance

This thesis study uses the MATLAB analysis of variance (ANOVA) function to evaluate the differences between groups of data. The null hypothesis (H0) tested by ANOVA is always the statement that the groups in question have the same mean. The p-value indicated by the ANOVA analysis indicates the strength of the null hypothesis and is compared against the stated confidence level termed alpha. Alpha determines how confident we are when rejecting the null hypothesis in the statistical analysis. Table 3.1 offers an interpretation of p-values with alpha set to 0.05 based on convention [7]. The table shows that the calculated p-value must be equal to or less than alpha in order to reject the null hypothesis with a 95%, ($(1-\alpha)\times 100\%$), confidence level.

P-value	Interpretation
$p<0.01$	Very strong evidence against H0
$0.01 \le p < 0.05$	Moderate evidence against H0
$0.05 \le p < 0.10$	Suggestive evidence against H0
$0.10 \le p$	Little or no real evidence against H0

Table 3.1 – Interpretation of the p-value used when analyzing the significance of acoustic cues in Chapters 4, 5, and 6.

As a part of the statistical analysis we also show the box plots of values extracted for the nasal and non-nasal groups of sonorant landmarks. Here we give a brief interpretation of the box plots from the MATLAB reference index. The lower and upper lines of the "box" are the 25th and 75th percentiles of the sample, while the horizontal line within the box (at the center or close to the center) is the sample median. Distance between the median line and the center of the box, is an indication of skewness of the

distribution, while the distance between the top and bottom of the box is called the interquartile range. Plus sign(s) that may appear outside the interquartile range are outliers in the data; by default, an outlier is a value that is more than 1.5 times the interquartile range away from the top or bottom of the box. For further interpretation of the box plots the reader should refer to the MATLAB help index or reference books.

We conclude this chapter by explaining the initial processing of the signal that is common to all algorithms explained and used in the next three chapters.

3.5 General processing of the signal

The analyses for all algorithms are conducted with short-time processing techniques commonly used in the processing of speech signals. The underlying assumption in the use of these techniques is that speech is quasi-stationary and its characteristics change slowly with time due to the physiological properties of the vocal tract. These properties allow short portions of the speech signal to be isolated and processed as if they belonged to a sustained sound. In order to produce such short segments, the speech signal x[n] is multiplied by a finite-duration window w[n] to produce snippets

$$s = w[n] \times x[n]. \quad (3.1)$$

For further characteristics and discussion on windowing, the reader can refer to any speech-processing textbook. As in many digital speech-processing applications, the window used in this thesis is a 25.6 ms hamming window applied every 10 ms. In the remaining chapters we will refer to the portion of the signal under a window as a frame.

In particular, we refer to "frame at time t" as the portion of the signal captured by placing the center of a 25.6 hamming window at time t. With a 16 kHz sampling frequency, each 25.6 ms window has 413 points allowing the algorithms to use a 512- or 1024-point DFT in calculating the short-time spectra. For algorithms that require calculation and analysis of the short-time spectra, we use the 1024-point DFT to enhance the frequency resolution to 15.75 Hz. All data analyses of short-time spectra are based on acoustic values that have been converted to decibels, using the equation

$$V_{db} = 20 \cdot \log_{10} V \text{ . (3.2)}$$

3.6 Summary

In this chapter, we explain the meaning and significance of acoustic criteria for nasality as tests that determine whether the signal around the landmark shows evidence of nasal characteristics. Depending on the region of the signal around the landmark to which they apply, we separate the criteria for nasality into a criterion for the nasal boundary, nasal murmur, and nasalized vowel. In the next three chapters we evaluate the existing acoustic criteria and cues on the set of 250 nasal and non-nasal sonorant landmarks. The results are analyzed in terms of the past observations and using the ANOVA statistical analysis function in MATLAB.

Chapter 4

4.1 Nasal boundary

In Chapter 3 we concluded that true nasal sonorant landmarks in a VCV database separate the signal into a region belonging to a vowel segment and a portion with the nasal murmur characteristics. In this chapter we attempt to formulate an acoustic criterion that describes the transition between these two regions by measuring two acoustic cues across 250[1] nasal and non-nasal sonorant landmarks estimated in Chapter 2. Our expectation is that the analysis of the quantified cues will show their applicability and effectiveness in separating nasal from non-nasal tokens as a part of the automated nasal detection module. Selected cues and quantitative expectation regarding their behavior will be included in the final nasal detection module, and tested on 1483 pivots estimated in the VCV database. The significance of developing an acoustic criterion for the nasal boundary extends to spontaneous speech – in American English at least one segment adjacent to the nasal is a vowel and will have the boundary properties described in the next sections. In this chapter we:

1. Propose two acoustic cues that have been found characteristic of the VN/NV boundary in past studies, which mostly used hand-measured observations,
2. Quantify and analyze them on the set of 250 estimated nasal and non-nasal sonorant landmarks,
3. Select effective cues and quantified expectation regarding their behavior for inclusion in the automated nasal detection module.

[1] For the remainder of this chapter we do not use the single member of the 'in semivowel' group of estimated sonorant landmarks. It is our opinion that analyses on a single-member group would carry limited significance.

4.2 Overview of the selected acoustic cues

Transition between a vowel and nasal segment is an acoustically rich area that is only partially examined during the landmark estimation. Our main objection is that the four examined frequency bands, Bands 2-5 specified in Chapter 2, are treated as equally important during the sonorant landmark analysis though it is expected that most robust acoustic changes will for the most part occur in the first and second formant region, and sometimes extend to the third formant frequencies. In this section, we describe the theoretical basis for the two acoustic cues that further examine the acoustic changes across [s] landmarks, and our approach to measuring them with automated algorithms in MATLAB. The cues compare acoustic properties of the signal, such as the energy in a band or formant region, before and after the landmark point for some time interval when attempting to determine whether they show properties of the nasal boundary.

Energy difference in the 0-350 and 350-1000 Hz frequency bands

Theoretical basis	The low first formant typical of the nasal murmur around 250Hz suggests that a possible characterization of the nasal boundary is the difference between the energy in the 0-350Hz and 350-1000Hz frequency bands. Transition from the vowel to the nasal segment involves lowering the first formant to the 200-350Hz range and an introduction of a nasal zero in the second formant region. Our expectation is that the energy of these two bands will be directly affected by the movement and widening of the first formant in the nasal segment and the introduction of the nasal antiformant at higher frequencies. In terms of the energy of the two bands, the lower band energy (LE) is expected to increase as we move into the nasal murmur region and as the location of F1 shifts to the 200-350Hz range. At the same time, the upper band energy (UE) decreases due to the lower energy of F2 and higher formants. Difference in the energy ED, defined as $ED = LE - UE$, (4.1)

	is thus expected to become more positive when measured across the nasal closure. The opposite is true for the nasal release into the adjacent vowel. The significance of this acoustic cue is not in the absolute value of either LE or UE, but in the way their difference, ED, changes across the landmark point.
Quantitative form	Glass calculates the energy in a particular band by taking the dot product of the short-time spectra $X(e^{jw})$ with a frequency window, $Z(e^{jw})$ [8]. Using Parsevals relation for conservation of energy, he shows that this calculation is equivalent to producing the short-time energy via equation: $E_n = \sum_{m=-\infty}^{m=\infty} (x_{bp}[m]w[n-m])^2$, (4.2) where $x_{bp}[m]$ is the result of passing the signal through the corresponding bandpass filter. Glass claims that this acoustic cue has been shown to successfully separate nasals from semivowels [8].
Algorithm	The first step in measuring the energy difference between the two frequency bands is to design two filters with the Matlab Filter Design Toolbox. One filter is a low-pass windowing filter with a cutoff frequency of 350 Hz. The filter is designed with 300 taps to allow for attenuation of -18 dB at 400 Hz. The second is a bandpass windowing filter that passes the 350-1000 Hz frequency range. The cutoff rate of -18 dB at 1050 Hz for this filter also requires 300 taps. Both filters use the hamming window, and have a linear phase and constant phase delay of 150 samples. Frequency response of each filter is shown in Appendix C. After the filtering stage, the energy of each short-time frame is calculated using the equation: $E_n = \sum_{m=-\infty}^{m=\infty} (x_{bp}[m]w[n-m])^2$, (4.3) where $x_{bp}[m]$ is the filtered signal of the desired band. The center of the first hamming window is at 20ms before the landmark point and last at 20ms after the landmark point. The time increment between consecutive frames is 10ms, producing a total of five observations. The output of the extraction algorithm for each frame consists of

	three parameters: 1. *LE* - energy in the low, 0-350Hz frequency band, 2. *UE* - energy in the 350-1000Hz band, 3. *ED* = *LE* – *UE* - difference in the energy between the lower and upper frequency band. Difference between the ED values measured in the first and last frame is considered to be the net energy change across the boundary for this time interval, denoted as ΔED.

H1 across the landmark boundary

Theoretical basis	Another measure that is characteristic of the vowel-sonorant boundary is the change in the first harmonic across the landmark point [5]. Stevens notes that we expect little change in the amplitude and spectrum of the glottal source during the interval from the preceding vowel through the murmur into the following vowel, resulting in essentially no change in the amplitude of the first harmonic throughout this time interval [23].
Quantitative form	Chen measures H1 by observing the energy of the first peak in the short-time spectra across the landmark boundary. She also cites a significant disparity in the first harmonic energy change for the transition between a vowel and obstruent versus vowel and nasal segment, making it a potentially useful measure to distinguish these two classes of estimated sonorant landmarks.
Algorithm	The complexity of measuring the first harmonic energy lies in approximating the location of the fundamental frequency with automated measurements. The approach we take in measuring this value is to obtain the f0 track of the utterance using the COLEA[2] function with the 'autocorrelation' option. We choose 'autocorrelation' because initial analysis of the F0 tracks indicated that in some instances F0 values measured with the 'cepstrum' approach reflect what appears to be phonetic glottalization of the segment. The values measured with the 'autocorrelation' method for the same utterances did not show similar characteristics. The F0 track is read in as a two-column matrix, where the first column of each row is the time of the center of the window in ms and the second, estimated F0 for that window in Hz. By default, COLEA sets time to

[2] COLEA: A Matlab Software Tool for Speech Analysis is a subset of a COchLEA Implants Toolbox developed at University of Texas in Dallas.

	increment in 20ms intervals (first 20ms window is centered at 10ms, then 30ms, 50ms, etc.). Once the F0 track is converted into a matrix, the algorithm centers the first window at 20ms before the landmark point and finds all local maxima of the 1024-point DFT spectra. For example, if a sonorant landmark point is estimated at 258ms, the center of the first hamming window will be at 238ms. After determining seven largest peaks in the short-time spectra, we read off the F0 value at the time closest to 238ms from the F0 track. Experimental analysis indicated that choosing five or more largest peaks of signal's DFT magnitude always included the first harmonic. H1 value is then approximated as the energy of the peak closest to the estimated F0. The algorithm next centers the window at the landmark point at 258ms and 20ms past the landmark point, at 278ms, each time repeating the described steps. This produces three observations of the H1 value: at the landmark, and 20 ms before and after the landmark point. In an attempt to minimize the effect of changing speech intensity, we propose to measure the absolute value of the first difference between consecutive H1 values, $$\left\|\Delta H1\right\| = \left\| H_{1\_at\_landmark} - H_{1\_at-20ms}\right\| + \left\|H_{1\_at\_landmark} - H_{1\_at+20ms}\right\|. \quad (4.4)$$ Large |ΔH1| values indicate greater fluctuation in the energy of the first harmonic across the stated time interval.

4.3 Analysis of the ΔED acoustic cue

When describing the acoustic characteristics captured by the ED cue, we noted that change in the energy of the two bands will depend on whether the cue is measured across a nasal closure or release. Theoretical expectation is that the shift of the first formant to lower frequencies and weakening of the second formant at the nasal closure will cause the LE to increase and UE to decrease. By the same token, release of the nasal

segment to the adjacent vowel will see an immediate upward shift in the location of the first formant, resulting in lower LE and increasing UE values. Defined as

$$ED = LE - UE, \quad (4.1)$$

ED is expected to show a positive net change across the nasal closure and negative change across the nasal release for some time interval around the landmark. The absolute value of that change will depend on the type of vowel adjacent to the nasal segment and degree of coarticulation between the vowel and consonant segment. In the next two sections we separate the analyses of the ED values based on the sign of the sonorant landmark around which they were measured.

4.3.1 ED cue measured at nasal [-s] landmarks

Our initial analysis focused on measuring the ED cue across 31 nasal [-s] landmarks, with the goal of determining whether the net energy change calculated with automated algorithms conformed to the defined theoretical expectations. Figure 4.1 illustrates the means of ED values measured at 10ms increments starting 20ms before and ending 20ms past the landmark point. The middle point in the graph corresponds to the ED values measured by overlaying the 25.6ms window symmetrically over the landmark point. Vertical bars in the graph represent ±1 standard deviation from the mean.

Visual inspection of the graph indicates that the mean of the ED value changes by about 10dB when measured 20ms before and 20ms after the landmark point, aligning with the expectation that nasal closure will cause a positive net change in the ED cue. The large spread of the observed values confirms our hypothesis that absolute values of

LE, UE, or ED will vary significantly based on the vowel type, place of articulation for the nasal segment, and amount of coarticulation between the two.

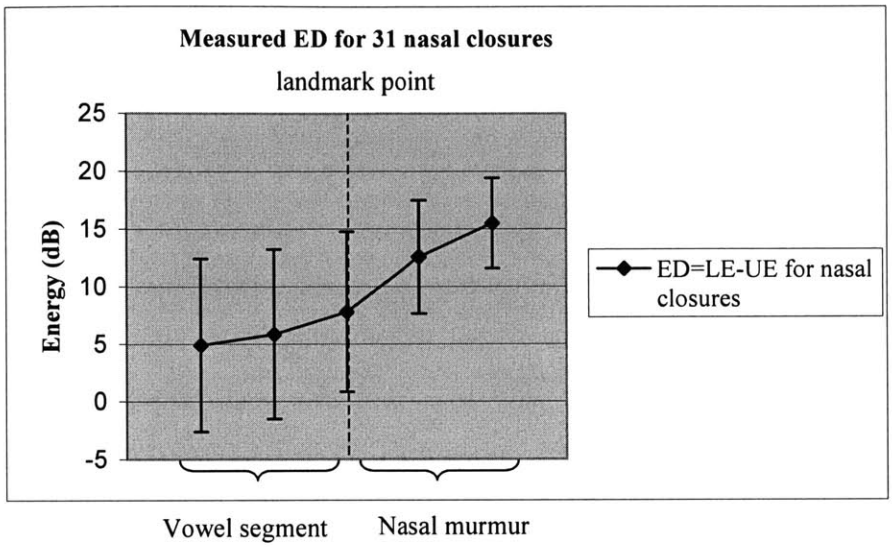


Figure 4.1 – ED values measured at different distances before and after 31 nasal [-s] landmarks. The positive net change agrees with the theoretical expectation that LE will increase and UE decrease across a nasal closure, for a positive net change in the ED value.

The net change for each nasal [-s] landmark estimated in the VCV database is a positive value between 2 and 25 dB that is somewhat dependent on the place of articulation for the nasal segment at the boundary. Figure 4.2 illustrates the net ED change for 31 nasal [-s] landmarks separated as measurements made across [m], [n], and [ŋ] boundaries.

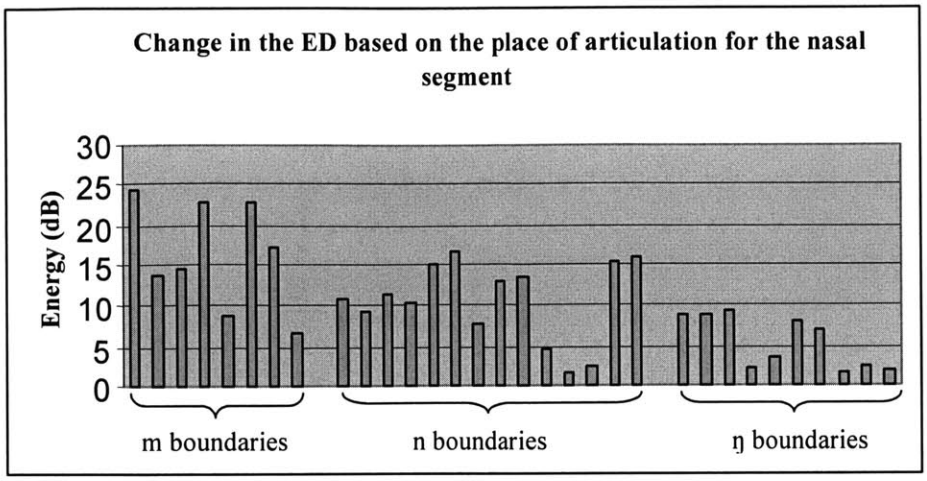


Figure 4.2 – Illustration of the individual net ED change for 31 nasal [-s] landmarks. The three groups correspond to landmarks at [m], [n], and [ŋ] boundaries for the three speakers.

4.3.2 ED cue around nasal [+s] landmarks

Repeating the same analysis on 39 nasal [+s] landmarks aligns with the stated theoretical expectation. Figure 4.3 illustrates the distribution of values for the ED cue measured at 10ms increments from 20ms before to 20ms past the nasal [+s] landmarks. The mean of the distribution changes by about -15 dB between the two most separated frames, for a total net negative change. The most abrupt change appears to coincide with the frame centered around the landmark point. The curve also seems to taper for the furthest points. Possible extension to the analysis of this cue would examine its behavior across a wider time interval and a possible expansion of the 350-1000 Hz range to 350-2000Hz to include the introduction of the nasal antiformant between 800 and 2000Hz.

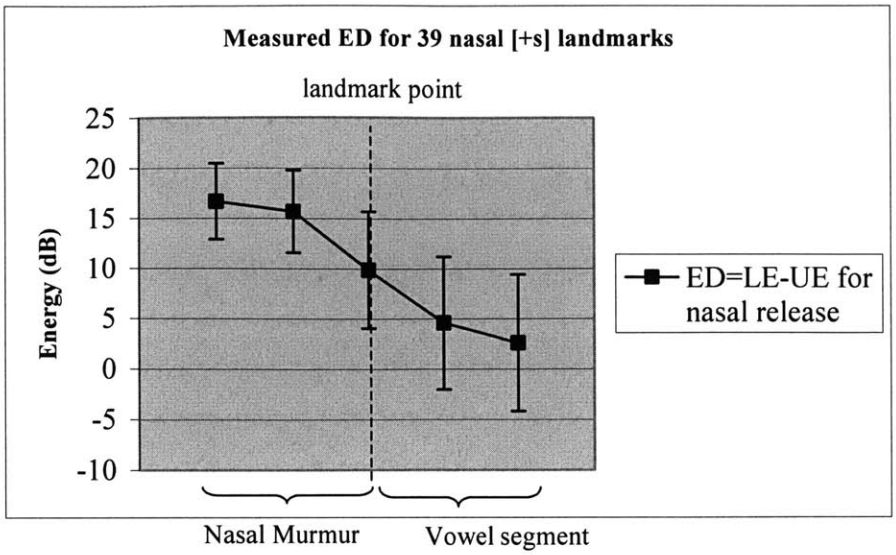


Figure 4.3 – ED value measured at different distances before and after a nasal [-s] landmark for 39 nasal [-s] landmarks. The utterances represent all three speakers. The positive net change for this time interval agrees with the theoretical expectation that LE will decrease and UE increase across the nasal closure.

Specific values measured for the nasal tokens show that each nasal release had a negative net change in the ED (ΔED) value that ranged from –5 to –25 dB for this time interval. The net change in the energy appears to be somewhat dependent on nasal's place of articulation as illustrated in Figure 4.4. The three groups of ED values are observations across the [m], [n], and [ŋ] boundary. The net energy change across a nasal release is on average larger than the change across a nasal closure, possibly reflecting the asymmetry of the coarticulation for the two types of nasal boundary. Characteristics of the vowel preceding the nasal segment reflect lowered soft palate for a longer time interval next to the boundary than the vowel following a nasal release.

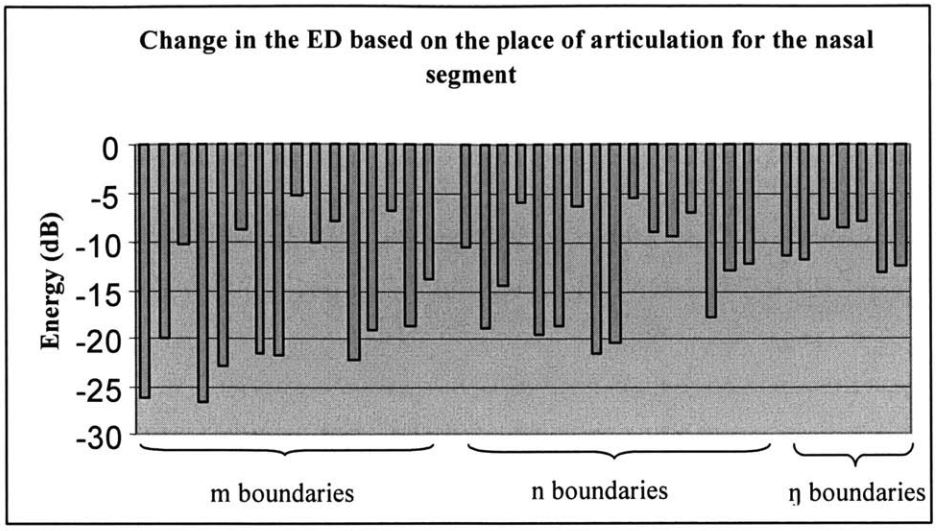


Figure 4.4 – Illustration of the individual net change in the ED value across 39 nasal [+s] landmarks, separated to [m], [n], and [ŋ] boundaries.

4.3.3 Applicability of the cue in detecting nasal boundaries

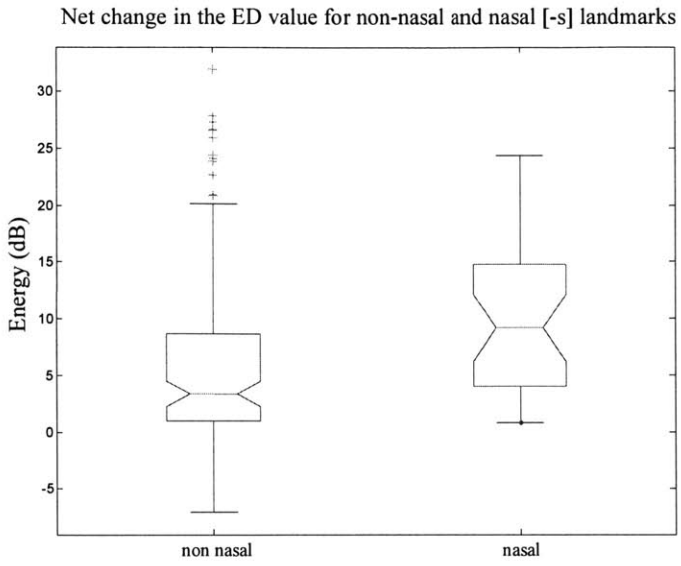
After verifying that measured values align with the theoretical expectation for the nasal landmarks, we examine the potential applicability of this cue in separating nasal from non-nasal sonorant landmarks estimated in Chapter 2. Comparison between the distributions of nasal and four non-nasal groups based on the landmark sign is greatly hindered by the small representation of some groups of estimated sonorant landmarks. The vowel-lateral/lateral-vowel boundary group, for example, has a total of 14 tokens, 3 of which are [-s] and 11 [+s] landmarks. Comparison between the nasal and such a small group of non-nasal sonorant landmarks would carry limited significance. Our approach to examining the applicability of this cue to the nasal boundary detection is to analyze the distribution of values for:

1. Nasal [-s] landmarks compared to the aggregate non-nasal [-s] landmarks,
2. Nasal [+s] landmarks compared to the aggregate non-nasal [+s] landmarks.

Final performance indication for this acoustic criterion is postponed for Chapter 7, when a greater number of pivots will determine its effectiveness in separating nasal from each group of non-nasal pivots.

4.3.4 Nasal versus non-nasal [-s] landmarks

Figure 4.5 illustrates the distribution of values for the net change in the ED value across this time interval for 31 nasal and 124 non-nasal [-s] landmarks. The median value for the nasal group is around 8dB, while the median for the non-nasal sonorant landmarks falls around 3 dB. The means of the two distributions differ by about 4dB. The ANOVA statistical analysis shows that the distributions of two groups are significantly different in terms of the observed mean and spread. With alpha set to 0.05, indicating the required 95% confidence level, a p-value of 0.011 indicates that the ΔED cue will have significantly different characteristics when measured at nasal and non-nasal [-s] landmarks for this time interval.



	Non-nasal	Nasal
Mean	6.16	10.27
St. Dev.	8.31	6.83

Figure 4.5 – Net change in the ED value for 154 nasal and non-nasal [-s] landmarks. The separation of distributions and means indicates that this acoustic cue is useful in detecting nasal boundaries for sonorant landmarks.

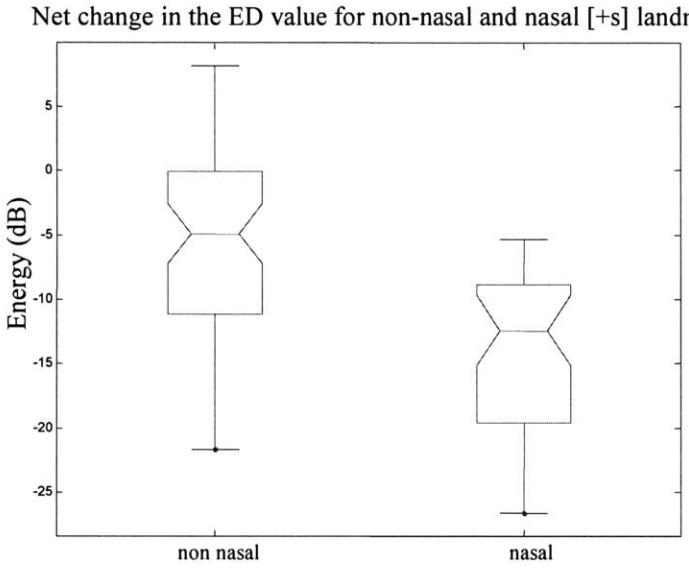
Table 4.1 illustrates the number of tokens analyzed for each group and an interpretation of the p-value. Because of the small number of tokens in the sub-classification of non-nasal landmarks, we do not attempt to test the effectiveness of this cue in separating nasal from each group of non-nasal sonorant landmarks.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Non-nasal	31	123	0.011	**Reject H_0**

Table 4.1 – Analysis of the two distributions with the ANOVA statistical analysis tool indicates that measuring ΔED at nasal versus non-nasal [-s] landmarks will yield significantly different values.

4.3.4 Nasal versus non-nasal [+s] landmarks

Repeating the same analysis steps for the [+s] landmarks shows that the ΔED cue will also produce significantly different values for nasal and non-nasal [+s] landmarks. In Figure 4.6 we illustrate the observed net change for 39 nasal and 57 non-nasal [+s] landmarks. Visual inspection of the graph shows that nasal [+s] landmarks produce negative values for the ED cue, with a median value of about –13dB. The range of values for the non-nasal group is mostly negative, with a median value of about –5dB. The means of two distributions differ by about 8dB.



	Non-nasal	Nasal
Mean	-6.07	-13.94
St. Dev.	7.16	6.29

Figure 4.6 – Distribution of ΔED values for 57 non-nasal and 39 nasal [+s] landmarks. The net change in the energy across nasal release for these tokens always falls below zero.

Next we analyze the distribution of values for the nasal and non-nasal group of [+s] landmarks using the ANOVA statistical analysis function and setting alpha to 0.05.

A p-value of 3.374×10^{-7} confirms that the distributions are significantly different in terms of the mean and spread. Qualitatively, a p-value below 0.05 means that measuring the ΔED acoustic cue across nasal and non-nasal [+s] landmarks will produce significantly different results. Table 4.2 illustrates the details of the ANOVA analysis and interpretation of the observed p-value.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Non-nasal	39	57	3.374×10^{-7}	**Reject H_0**

Table 4.2 – Analysis of the two distributions with the ANOVA statistical analysis tool indicates that measuring ΔED at nasal and non-nasal [+s] landmarks will yield significantly different values.

4.3.5 Formulation of an acoustic criterion based on the ΔED cue

Analysis in the previous section indicated that ΔED assumed strictly positive values across a nasal closure and strictly negative values across the nasal release when measured 20ms before and after the landmark point in the VCV database. Based on this observation we require that the ΔED acoustic cue shows a positive net change across [-s] and negative net change across [+s] landmarks for the same time interval when claiming that a given sonorant landmarks shows characteristics of the vowel-nasal boundary. A tighter criterion is contingent on further analysis of other databases and contexts.

4.4 Analysis of H1 across the nasal boundary

Unlike the energy change cue described in the previous section, H1 is expected to remain relatively unchanged across both types of nasal boundary – nasal closure and release – in a VCV database. As a reminder to the reader, we propose to measure the absolute value of the first difference in H1 between consecutive time frames starting

20ms before and ending 20ms past the sonorant landmark point. Center of each consecutive frame is incremented by 20ms. In the remainder of this section we will refer to this value as |ΔH1|, where

$$|\Delta H1| = \left| H_{1at\_landmark} - H_{1at-20ms} \right| + \left| H_{1at\_landmark} - H_{1at+20ms} \right|. \quad (4.4)$$

With 250 of nasal and non-nasal sonorant landmarks classified across five groups in Chapter 2, we analyze the values in two ways:

1. We first compare the measured values against the theoretical expectation to verify that the automated algorithms are functioning correctly,
2. Using the ANOVA statistical analysis tool we determine how |ΔH1| values measured across nasal sonorant landmarks differ from those measured for each non-nasal group.

4.4.1 Fluctuation of the |ΔH1| value across nasal sonorant landmarks

With a mean value of 1.60dB and standard deviation of 1.03dB for 70 nasal sonorant landmarks, the energy of the first harmonic does not appear to change significantly across the nasal sonorant landmark. This observation is qualitatively in agreement with Chen [5]. Figure 4.7 illustrates the |ΔH1| values measured for 70 nasal sonorant landmarks.

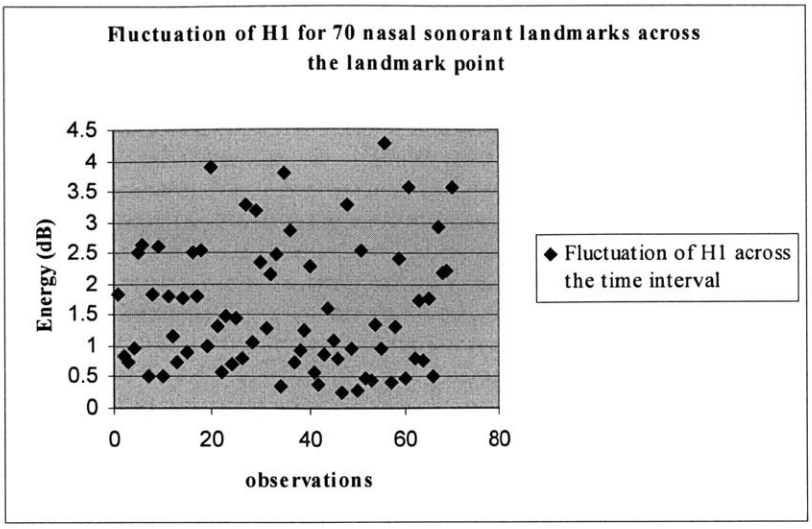


Figure 4.7 – Observation of the |ΔH1| cue for 70 nasal sonorant landmarks in the time interval starting 20ms before and ending 20ms after the landmark point.

Fluctuation in the |ΔH1| values can potentially be attributed to the automated estimation of the pitch performed by the COLEA software tool – information regarding its accuracy does not appear to be available from the developers' website[3]. In addition, COLEA determines the pitch at set times that may differ up to 10ms from the time of the peak detection – it is unclear whether this time misalignment is significant enough to contribute to the fluctuation in the energy of the first harmonic.

4.4.2 |ΔH1| cue measured across nasal and non-nasal sonorant landmarks

In this section we compare |ΔH1| values observed across the nasal and each non-nasal group of sonorant landmarks. Similarities between nasal, lateral, and semivowel groups of sonorant landmarks in Figure 4.8 align with the theoretical expectation; with negligible increase in the pressure above the glottis and no change in the spectrum of the glottal source during the interval from the preceding vowel through the sonorant segment

[3] http://www.utdallas.edu/~loizou/speech/colea.htm

and into the following vowel, sonorant boundaries show essentially no change in the amplitude of the first harmonic [25].

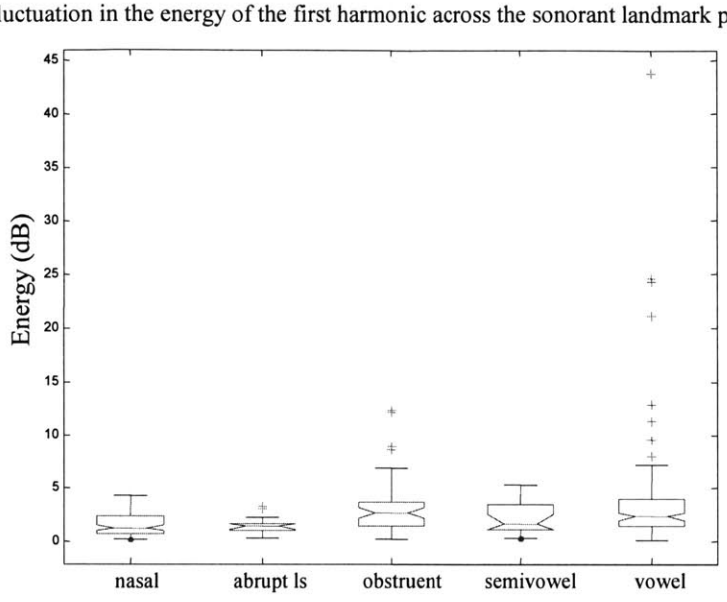


Figure 4.8 – Energy of the first harmonic measured across 250 nasal and non-nasal sonorant landmarks. The graph illustrates similarities between the nasal, lateral, and semivowel boundaries, for which negligible increase in the pressure above the glottis results in the continuity of the low-frequency spectrum amplitude.

Sonorant landmarks inserted within vowel segments surprisingly show the largest fluctuation in H_1; analysis of the specific instances that cause outlier values in Figure 4.8 indicates that these points represent sonorant landmarks at the beginning or end of the utterance. These contexts appear to cause a large fluctuation in the energy of the first harmonic. The small change in H_1 for the obstruent group of sonorant landmarks is possibly caused by the significant low-frequency energy that was observed for this database. Many of the vowel-obstruent boundaries for voiceless segments appear to be voiced for some time interval in the consonant segment due to this noisy component. Elimination of the observed noise would likely boost the effectiveness of this cue in

separating nasal landmarks from those inserted at obstruent boundaries. We next use the ANOVA statistical analysis tool with alpha set to 0.05 to determine this cue's applicability in the nasal boundary detection.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Abrupt [l]	70	14	0.853	**Cannot reject H_0**
Nasal	Obstruent	70	50	1.903×10^{-6}	**Reject H_0**
Nasal	Semivowel	70	18	0.007	**Reject H_0**
Nasal	Vowel	70	98	0.001	**Reject H_0**

Table 4.3 – Details of the statistical analysis of the |ΔH1| cue confirm the similarity between distributions of sonorant landmarks at nasal and lateral boundaries. Although not anticipated, the |ΔH1| cue appears to show different characteristics when measured at nasal and semivowel boundaries. Nasal boundaries also differ from sonorant landmarks inserted at obstruent boundaries and within vowel segments.

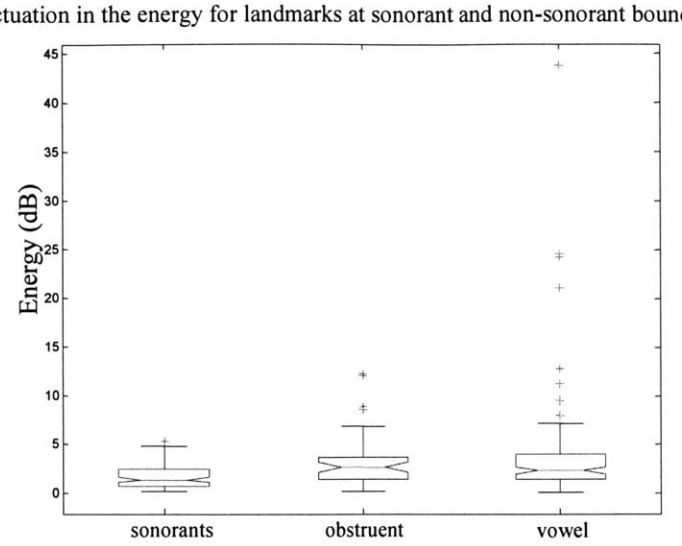
Statistical analysis results in Table 4.3 confirm the similarity between sonorant landmarks at nasal and lateral boundaries, and variation in the values for the remaining groups. ANOVA statistical analysis surprisingly produces a relatively high (above alpha) p-value for the nasal and semivowel groups of estimated sonorant landmarks, indicating that their distributions are significantly different. It is unclear as to why semivowels do not show the same properties as the remaining sonorant groups. The lowest p-value, indicating greatest dissimilarity, is calculated in a pair-wise statistical analysis between nasals and obstruents, which is in agreement with Chen's observation [5].

4.4.3 Formulating an acoustic criterion based on the |ΔH1| cue

An approach to formulating an acoustic criterion based on the |ΔH1| cue relies on the following two observations:

1. Acoustic criterion should be formulated in terms of the largest permissible value for |ΔH1| in the stated time interval starting 20ms before and ending 20ms after the landmark point – this requirement aligns with the stated theoretical basis that requires relatively constoant energy of the first harmonic across vowel-sonorant boundaries,
2. The criterion should include values for the nasal, lateral, and semivowel groups of estimated sonorant landmarks, because they are expected to show the same acoustic characteristics at low frequencies.

Figure 4.9 illustrates the distribution of |ΔH1| values for landmarks at true sonorant boundaries (nasals, laterals, and semivowels), and vowels and obstruents. With the mean value of 1.72 and standard deviation of 1.17 for the aggregate distribution, we propose to allow |ΔH1| values to fall within three standard deviations from the mean, or up to 5.23dB. Chapter 7 further examines this acoustic cue by analyzing the performance of the stipulated acoustic criterion on 1483 pivots – we are mainly concerned with the performance of COLEA's pitch detection algorithm in different contexts and possible variability in its performance.



	Sonorants	Obstruents	Vowel
Mean	1.72	3.37	3.97
St. Dev.	1.17	2.72	5.85

Figure 4.9 – Sonorant group aggregates values measured for the nasal, lateral, and semivowel groups of estimated sonorant landmarks because they are expected to show the same characteristics at low frequencies.

4.5 Summary

In this chapter we set out to examine properties of the nasal boundary that are not captured during the sonorant landmark estimation. The two cues that we propose focus on the acoustic characteristics in the first and second formant frequency range. The goal is to select cues that can be implemented in terms of automated algorithms in MATLAB and that show distinctive characteristic when measured across nasal versus all other boundaries. The selected cues are included in the nasal detection module that is tested on the 1483 estimated pivots in Chapter 7.

The first cue that we described and tested on 250 nasal and non-nasal sonorant landmarks monitored the difference in the energy of the lower, 0-350Hz, and upper, 350-

1000Hz frequency bands. Values measured with automated algorithms conformed to the theoretical expectation and were found to produce significantly different values when measured at nasal and non-nasal sonorant landmarks. Observed ΔED mean for nasal and non-nasal landmarks differed by 4dB when measured across [-s] and 8 dB across [+s] landmarks.

The next described cue was based on the expectation that sonorant boundaries will see negligible change in the energy of the first harmonic due to the small change in the pressure at the glottis. The observed |ΔH1| values were in relative agreement with the theoretical hypothesis and showed to separate true sonorant landmarks from landmarks inserted at obstruent boundaries and at the beginning or end of the utterance.

At the end of this chapter we select the two cues, |ΔH1| and ΔED to include in the final design of the nasal detection module in Chapter 7.

Chapter 5

5.1 Nasal Murmur

In this chapter we discuss and formulate an acoustic criterion for the nasal murmur by examining and modifying Chen's algorithm for nasal murmur detection. In the original study, this algorithm used six hand-measured acoustic cues to correctly reject 54.5% of non-nasal landmarks and confirm all examined nasal sonorant landmarks for the nasal murmur characteristics. Despite the low overall non-nasal rejection rate in the original study, the algorithm discriminated very well against the sonorant landmarks within vowel segments and at vowel-semivowel boundaries, with rejection rates of 87.5% and 64% respectively. The reason why we choose to examine and possibly improve this algorithm for the nasal murmur is that its structure lends itself to implementation in terms of automated measurements. It is also expected that the low rejection rate will be partially offset by the acoustic criteria for the nasal boundary and nasalized vowel. In this chapter we:

- Flesh out Chen's algorithm in terms of automated measurements in MATLAB,
- Examine its performance on the 250 hand-classified nasal and non-nasal sonorant landmarks from Chapter 2,
- Suggest possible improvements to the algorithm,
- Test the algorithm with the proposed modifications on the same set of sonorant landmarks and discuss its performance.

5.2 Overview of the acoustic cues in Chen's algorithm

As discussed in Chapter 3, each sonorant landmark separates the signal into a region with a higher and lower energy level. Chen's algorithm for the nasal murmur analyzes the portion of the signal with a lower energy level by extracting six acoustic cues that describe the energy and spectral tilt characteristics. The algorithm then compares the observed values against the formulated expectation, and rejects some sonorant landmarks as not showing properties of the nasal murmur. In this section we introduce the six acoustic cues used in Chen's algorithm. Information regarding each cue is presented as a three-entry table:

Theoretical basis	The first entry gives a brief overview of the theoretical basis for why the acoustic cue is characteristic of the nasal murmur.
Quantitative form	In the second entry we include the quantified form as presented by Chen, and the cue's effectiveness in separating nasal from non-nasal landmarks where available.
Algorithm	This entry describes the implementation in terms of automated algorithms in MATLAB used for the present study.

Energy characteristics of the signal

Theoretical basis	Holding the vocal apparatus fixed during nasal consonant production raises the question of how much do the energy parameters change during the nasal murmur. With little change in the vocal tract configuration, the energy in this signal portion should be relatively constant.
Quantitative form	The energy stability criterion proposed by Chen is measured by calculating the first difference between the RMS value of consecutive window frames. Chen requires that the change in the RMS value between two consecutive frames does not exceed 1 dB for a nasal murmur. The information regarding this parameter's effectiveness is unknown.

Algorithm	The short-time RMS value is defined as, $RMS_{frame} = \sqrt{(x[m] \times w[m])^2}$, (5.1) where w[m] is a 25.6 ms Hamming window applied every 10 ms. The algorithm calculates the first difference in the RMS value of consecutive time frames as $\Delta RMS = \left\| RMS_{frame+1} - RMS_{frame} \right\|$ (5.2) and compares this value against a firm threshold.

Low-frequency energy centroid f1

Theoretical basis	Numerous studies of nasal consonants note a characteristic spectral shape of the nasal spectrum, dominated by a low frequency prominence between 200 and 350 Hz. To a first approximation, this pole is a Helmholtz resonance between the acoustic compliance of the vocal tract volume and the acoustic mass of the nasal passages.
Quantitative form	Chen suggests locating the low-frequency energy centroid by identifying the highest energy peak in the 0-788 Hz frequency range. For a nasal segment, Chen expects this value to fall between 126 and 347 Hz. For fricative obstruent segments, for example, this value will most likely fall at the upper limit of the proposed range. Glass claims that the presence of this low frequency prominence is a necessary, but not sufficient, condition for the identification of the nasal murmur.
Algorithm	The frequency of the low-frequency energy centroid f1 is evaluated by first calculating the short-time frequency spectra of the windowed signal using a 1024-point DFT. Within the short-time spectrum, the peak picking function identifies each of the local maxima in the 0-788 Hz frequency band, and selects the largest peak. The frequency of the largest peak is considered to correspond to the expected prominence, termed f1. The peak-picking function has no requirements in terms of the absolute or relative peak height.

Spectral Tilt

Theoretical basis	Three factors contribute to the characteristic spectral tilt of the nasal murmur. The first contributing factor is the low frequency of the first pole, which tends to reduce the amplitudes of the remaining poles in the transfer function. With a weak second pole of the transfer function in the range of 750 to 1000 Hz and an antiformant due to the nasal cavity in the 800-2000 Hz range, the spectral shape is tilted toward low frequencies. The overall shape typically involves high energy at low frequencies and a rapid drop at frequencies above 1000 Hz.
Quantitative form	In trying to capture the spectral tilt, Chen suggests measuring the relative differences between the largest spectral prominences in five frequency bands. The frequency bands are 0-788, 788-2000, 2000-3000, 3000-4000, and 4000-5000 Hz. If we denote the largest resonance in each frequency band as A_1 through A_5, the suggested parameters are A_1-A_2, A_1-A_3, A_2-A_3 and the Sum of Amplitude Differences (SAD), where SAD = $(A_1-A_2)+(A_1-A_3)+(A_2-A_3)+(A_1-A_4)+(A_1-A_5)$. Measuring the *difference* between the resonances is especially suitable because the energy of each resonance does not need to be calibrated for speech intensity. For nasal murmur, Chen expects the following ranges for each parameter: $11.8<A_1-A_2<43.3$ dB $22.8<A_1-A_3<50.3$ dB $-9.7<A_2-A_3<33.6$ dB $119.5<SAD<200.7$ dB
Algorithm	The algorithm used to measure this set of acoustic parameters is very similar to the one used to determine the location of the energy centroid. The short-spectra are divided into five frequency bands, as recommended by Chen. Within each band, a peak picking function identifies all local maxima and selects the one with the highest energy. The selected peaks are denoted A_1 through A_5, where the subscript indicates each peak's frequency band. In the final step, the suggested parameters, A_1-A_2, A_1-A_3, A_2-A_3, and SAD, within each frame are constructed using these values.

In the next section we describe the way the six cues are used in the reconstruction of Chen's algorithm.

5.3 Reconstruction of Chen's algorithm for the nasal murmur detection

When measuring nasal murmur cues, the algorithm places the center of the first 25.6ms window at 30ms past the [-s] and before a [+s] landmark time. The goal of excluding a time interval immediately surrounding the estimated landmark time is twofold:

1. Not including the acoustic characteristics of the signal surrounding the landmark time ensures that the nasal murmur cues are not influenced by the abrupt landmark boundary,
2. Slight temporal misalignment of the true and estimated landmark time will not influence the observed acoustic cues
    a. This approach allows for the possibility that true landmarks fall 15ms past the estimated landmark point, without affecting the measurements,
    b. For true landmarks that fall before an estimated landmark this approach will shorten the region corresponding to the nasal murmur by starting measurements even further away from the true landmark.

Figure 5.1 illustrates the algorithm estimation. The first window or Frame 1, centered at 30ms past the landmark point produces an observation of only one acoustic cue – the signal's RMS value. This RMS value is denoted as RMS_1 and saved for comparison with the RMS value of the next frame. The window then moves 10ms further into the region with a lower energy level (to the right for a [-s] and left for a [+s] landmark), and measures all six acoustic cues, RMS_2, f1, A_1-A_2, A_1-A_3, A_2-A_3, and SAD. The center of

the window is now at 40ms past the landmark point[1] or Frame 2 as referred to in the rest of this section.

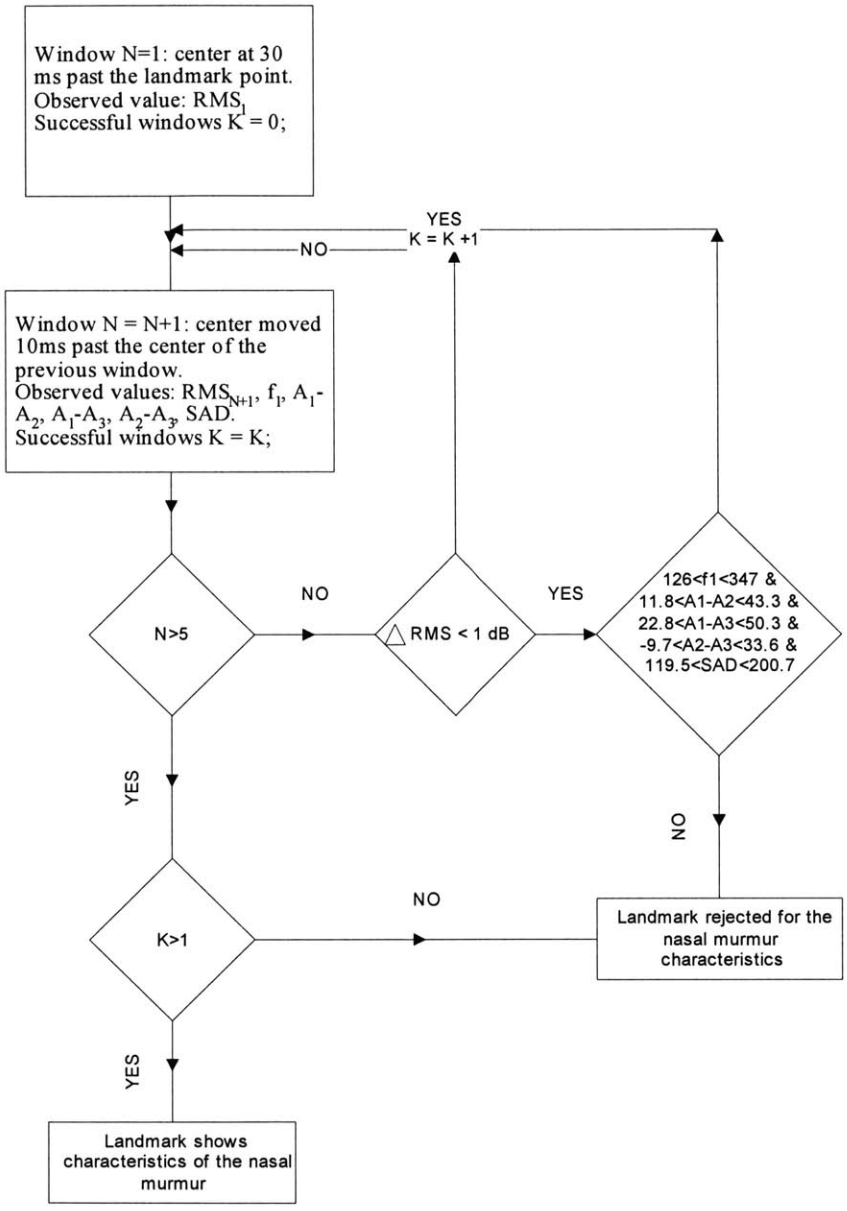


Figure 5.1 - Illustration of the reconstructed algorithm for the nasal murmur detection proposed by Chen. The algorithm examines at most five frames around the landmark time for the nasal murmur characteristics, and expects at least two to pass all acoustic criteria to consider the landmark as showing characteristics of the nasal murmur.

[1] By time frame at time x, we refer to that portion of the signal captured by placing the center of the 25.6ms Hamming window at time x. In Chen's algorithm we term the portion of the signal captured by placing the center of the first window at 30ms past the landmark point Frame 1 and increment the count each time we slide the window by another 10ms.

The algorithm calculates ΔRMS as the absolute difference between RMS_2 and RMS_1, and proceeds to compare this value against the RMS criterion. If ΔRMS is less than 1dB, the algorithm compares the five remaining cues against the expectation formulated in terms of the acoustic criteria. A single cue that falls outside its permissible range causes the sonorant landmark to be rejected for the nasal murmur characteristics. If all five cues fall within the permissible range, however, the algorithm slides the window by 10ms in the same direction, and repeats the process for Frame 3 centered at 50ms past the landmark point.

If the difference between the two RMS values is greater than 1dB, the frame is discarded and the algorithm slides the window by another 10ms to examine the next set of acoustic cues. The RMS value of the discarded frame still becomes the new reference for comparison with the next frame. Chen continues to verify the acoustic criteria for as long as the RMS value of consecutive frames does not change by more than 1 dB. The goal of using the change in the RMS value to determine whether a frame should be examined for the remaining acoustic cues is to ensure that measurements are made in the portion of the signal that shows the steady-state energy characteristic of the nasal murmur.

Chen gives no precise definition of the maximum number of frames examined by the algorithm. In this implementation and evaluation of the algorithm, we make an assumption that no more than five frames need to be examined for the nasal murmur. Five frames are chosen as reasonable because the center of Frame 5 is 70ms past the landmark point, and the characteristics of the signal are no longer expected to adhere to the expectation for the original landmark type. Because Frame 1 gives only the RMS reference point, five frames give a total of four observations of the six acoustic cues in

Frames 2-5. Another assumption that we make is that a landmark showing properties of the nasal murmur will have at least two of four examined frames pass all acoustic criteria. The duration requirement implicit in this assumption is that nasal murmur extends at least 63ms past the landmark point (for the case when Frames 2 and 3, centered around 40 and 50ms respectively, pass all acoustic criteria and the remaining frames do not meet the ΔRMS requirement). Analysis of the VCV database shows that the signal around all nasal sonorant landmarks from Chapter 2 shows characteristics of the nasal murmur for at least 80ms past each estimated landmark point, well beyond the requirement. Discussion on the application of this algorithm to spontaneous speech, where nasal murmur can be shorter than 50 ms as suggested in Chen is reserved for Chapter 7.

5.4 Performance of Chen's algorithm for the nasal murmur detection

The original algorithm proposed by Chen appears to impose a rigid set of acoustic criteria that each sonorant landmarks needs to meet for the nasal murmur characteristics. A consequence of this structure is a high rejection rate across all classes of estimated sonorant landmarks as defined in Table 2.3. If we define the rejection rate for each of the five classes as:

$$rejectionRate = \frac{number\_of\_rejected\_tokens}{number\_of\_tokens\_in\_class} * 100\%, \quad (5.3)$$

the observed rejection rates for the automated nasal murmur algorithm are significantly higher than those reported in Chen.

The rejection rates for non-nasal sonorant landmarks vary from 89.1% for landmarks inserted within vowel segments to 71.4% for landmarks at abrupt [l] boundaries. The average rejection rate across the non-nasal groups of sonorant landmarks is 85.6%. Table 5.1 shows the number of tokens tested within each class, number of tokens rejected for the nasal murmur, and the rejection rates for each of the five classes of estimated sonorant landmarks.

Class of Sonorant Landmark	Comment	Total [s] landmarks	Landmarks rejected for nasality	Rejection Rate
Nasal	True positive	70	24	34.3%
Abrupt l boundary		14	10	71.4%
Within vowel	False positive	98	90	91.8%
SV/VS boundary		18	14	77.8%
OV/VO boundary		50	41	82%
Total	**Nasal**	70	24	34.3%
	Non-nasal	180	155	86.1%

Table 5.1 – Implementation of Chen's algorithm in MATLAB with the original set of acoustic criteria. Although efficient in terms of rejecting non-nasal sonorant landmarks, the algorithm performance is offset by a relatively high error rate of 34.3% for nasal sonorant landmarks.

The efficiency of the high rejection rate for non-nasal sonorant landmarks is partially offset by a significant error rate; with the proposed acoustic criteria Chen's algorithm also discards 34.3% of all nasal sonorant landmarks, mainly due to failures to meet the criteria for A_1-A_3 and SAD[2].

Difference in the rejection rate for non-nasal sonorant landmarks in Chen and here can be attributed to the difference in the structure of the two databases. As previously

[2] Examination of the VCV database indicated that rejected nasal landmarks did exhibit characteristics of the nasal murmur for at least 80ms past the landmark point.

mentioned, it is expected that stressed syllables in a VCV database are articulated with greater care, possibly resulting in a cleaner, more robust acoustic signal and better performance of the original algorithm[3]. Chen's study, however, cites no instances where the proposed algorithm erroneously rejected nasal sonorant landmarks for nasal murmur as observed here. This performance aspect is possibly caused by:

1. A discrepancy between the values obtained through hand-measurements in Chen and automated algorithms in this study,
2. Chen's decision to formulate an acoustic criterion for nasal murmur that is finely tuned to accept all nasals in the original database, rather than being general enough for other speakers and contexts.

We thus turn our attention to the analysis of the specific values measured for the six acoustic cues as a means of identifying the cause of nasal rejection. Our goal is to examine the distribution of measured values and suggest modifications that will minimize the number of nasal sonorant landmarks rejected for the nasal murmur without significantly affecting the non-nasal rejection rate.

5.5 Evaluation of Chen's acoustic cues and their effectiveness

As a reminder to the reader, Chen's algorithm for the nasal murmur detection measures the first acoustic cue, $RMS_{Frame\ 1}$, by centering the 25.6ms hamming window at 30ms past the landmark point. In the algorithm description and diagram, this frame is denoted as Frame 1 (n=1). Sliding the window by 10ms produces Frame 2, and the first

[3] Chen analyzed the performance of the algorithm on the LAFF database of grammatically correct sentences.

observation of the six acoustic cues[4]. In this section, we analyze the observed values for each acoustic cue in Frame 2 on 250 nasal and non-nasal sonorant landmarks from Chapter 2.

Analysis of each cue is divided in three sections.

1. In the first section, we analyze the distribution of values for a specific cue based on the landmark type around which it was extracted. Using the ANOVA statistical analysis function we examine whether distributions of nasal and each non-nasal group of sonorant landmarks are statistically different. This information allows us to estimate the effectiveness of the cue in separating nasal from non-nasal sonorant landmarks.
2. Next, we focus on the nasal group of estimated sonorant landmarks and compare the distribution of values to Chen's suggested criteria – results of this analysis determine the source of errors observed in the performance of the algorithm.
3. Lastly, we use our previous observation that each estimated nasal sonorant landmark from Chapter 2 shows acoustic characteristics of the nasal murmur for at least 80ms past the landmark point to determine how each acoustic cue changes when measured at different distances from the landmark point. The question we attempt to answer is – can we use the information obtained by analyzing the acoustic cues on a single frame to formulate an acoustic criterion for any nasal sonorant landmark and any frame?

The reason why we choose to focus the analysis on a single frame is the applicability of the algorithm to spontaneous speech. With no information regarding the duration of

[4] The RMS acoustic cue measures the relative change in the RMS value between consecutive frames – for this reason the first RMS value serves only as a reference and not an absolute acoustic cue.

the nasal segment in an automated speech recognition module, it may be important to formulate the decision process such that it relies on only one or two available frames. In addition, nasal murmur can be significantly shorter in spontaneous speech, often not exceeding 50ms in duration from the time of the nasal closure. With the current formulation of the algorithm, which places the center of the first window at 30ms past the landmark point, only one frame would be completely contained within the 50ms murmur region. Choosing to analyze the values of cues in the frame closest to the estimated landmark point follows the reasoning that the signal in the immediate vicinity of the landmark exhibits the strongest characteristics of the landmark type.

When examining how values of the six cues change as we move further away from the landmark point, we recall our previous observation that each estimated nasal sonorant landmark from Chapter 2 has nasal murmur that extends at least 80ms past the landmark point. With the windows centered at 40, 50, and 60ms past the nasal sonorant landmark point, we can examine the distribution of values in three frames (Frame 2, 3, and 4 respectively) and still remain in the nasal murmur region. The 25.6ms Hamming window centered at 60ms past the estimated landmark point is within the hand-measured duration of all sonorant segments in the VCV database.

5.5.1 RMS

5.5.1.1 Distribution of RMS values across landmark types

Figure 5.2 displays the range of values observed for ΔRMS between Frames 1 and 2 based on the type of sonorant landmark around which they were extracted. Small mean value and a tight distribution of the proposed RMS cue for the nasal group validate the expectation that the fixed vocal tract configuration characteristic of the nasal murmur

results in a virtually constant signal in terms of energy. The energy stability is to a large extent mirrored by the sonorant landmarks at lateral boundaries – this group of landmarks also shows small mean and fluctuation in the energy between Frames 1 and 2. Unlike these two groups, sonorant landmarks inserted within vowel segments appear to have the widest distribution of values and most outliers for the ΔRMS cue. The reason for the wide spread is the position of the estimated landmark in the vowel. Depending on where the estimated landmark is in the vowel, centering a 25.6ms hamming window at 30 and 40ms past the landmark point can include portions of the signal at the beginning or end of utterance, or even at the boundary of the vowel and consonant segment. These three events are often associated with rapidly changing energy of the signal.

A pair-wise comparison of the ΔRMS cue for the nasal and each group of non-nasal sonorant landmarks with the ANOVA statistical analysis function confirms the difference in the means and spreads for the five landmark groups. Setting alpha to 0.05, thus requiring a 95% confidence level, indicates that measuring the ΔRMS cue at nasal and all but sonorant landmarks at lateral boundaries will produce a statistically different distribution of values. For landmarks at lateral boundaries, distribution of ΔRMS values for Frame 1 and 2 is not considered significantly different from the nasal distribution.

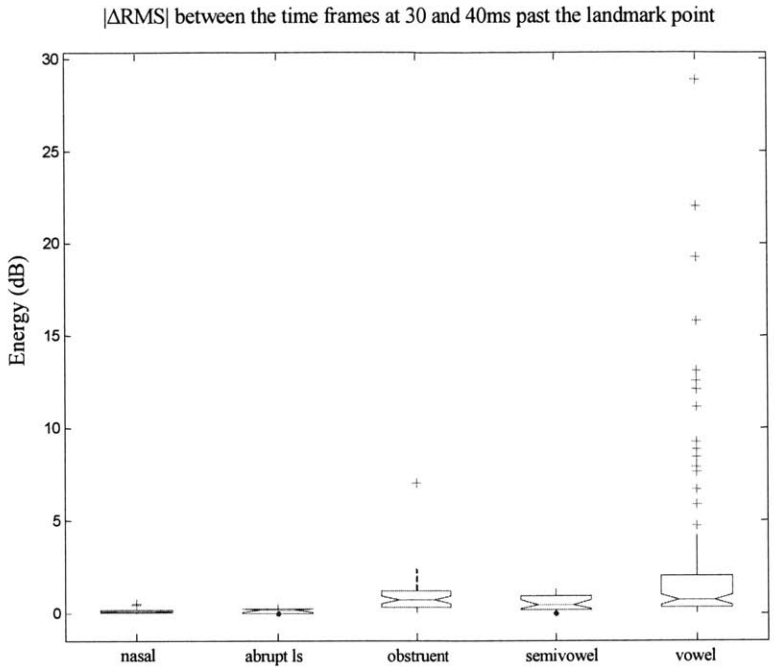


Figure 5.2 – Distribution of the ΔRMS values for the nasal and non-nasal groups of sonorant landmarks. The nasal group is characterized by the lowest mean and tightest distribution, followed by the laterals, semivowels, obstruents, and lastly landmarks inserted within vowel segments.

Table 5.3 shows the number of tokens in each group and the meaning of the calculated values. Based on the analysis so far, we conclude that ΔRMS is a significant acoustic cue that can be used to separate nasal from all but the lateral group of non-nasal sonorant landmarks.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Abrupt [l]	70	14	0.139	**Cannot reject H_0**
Nasal	Obstruent	70	50	1.533×10^{-8}	**Reject H_0**
Nasal	Semivowel	70	18	2.707×10^{-10}	**Reject H_0**
Nasal	Vowel	70	98	3.194×10^{-5}	**Reject H_0**

Table 5.2 – ANOVA analysis of the extracted RMS parameter between two consecutive frames centered at 30 and 40ms suggests that this cue has statistically different means when measured around nasal and non-nasal sonorant landmarks. Formulating the quantitative expectation for this cue's behavior appears to be a promising acoustic criterion for the nasal murmur detection.

5.5.1.2 Distribution of ΔRMS for the nasal group of estimated sonorant landmarks

Focusing on the distribution of ΔRMS values for the nasal group of sonorant landmarks indicates that the observed values are well within the range specified by Chen, as illustrated in Table 5.3. The difference between the two ranges appears to stem from the limitation of the Klatt speech software tool used in Chen's study. When calculating the RMS difference between two consecutive time frames, Klatt tool approximates the calculated value to the closest integer. This limits the available resolution in Chen's RMS measurements to 1 dB, which is significantly lower than the hundredth of decibel accuracy used in this study. The ΔRMS cue, thus, is not responsible for the erroneous rejection of nasal sonorant landmarks observed in the automated implementation of Chen's algorithm.

	Nasal	
	Automated	Chen
Min.	0.00	0.00
Max.	0.53	1.00
Mean	0.14	n/a
St.Dev.	0.12	n/a

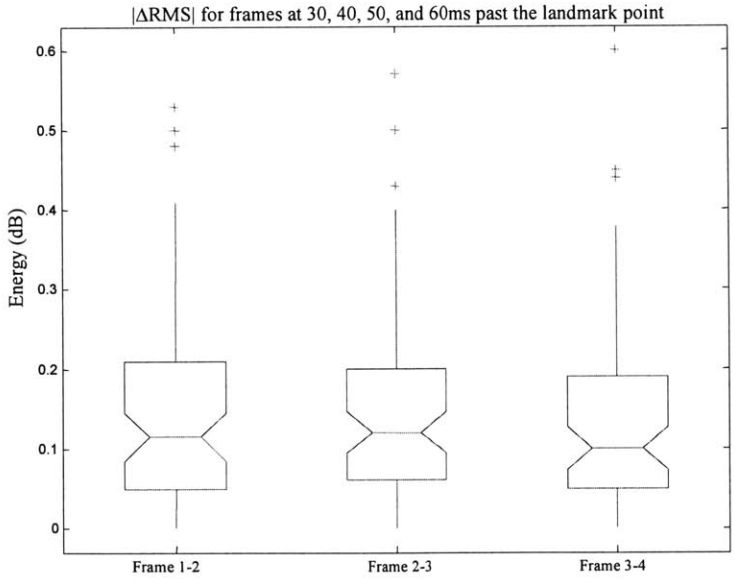
Table 5.3 – Overview of the RMS values extracted with automated algorithms and those suggested in Chen. The difference in the range of values stems from the limitation of the Klatt software tool used in Chen's study – by approximating each RMS value up to the next integer, the Klatt tool limits the available precision to 1 dB compared to hundredths of a decibel used in this study.

5.5.1.3 ΔRMS cue as a function of distance from the landmark point

Lastly, we turn our attention to the behavior of ΔRMS when measured in nasal murmur, but at different distances from the landmark point. Our earlier observation that each nasal sonorant landmark in the VCV database has nasal murmur that extends at least 80ms past the landmark point allows us to obtain three measurements of the RMS cue:

1. First value is the change in the difference of the RMS value between the time frames centered at 30 and 40ms, denoted as Frame 1-2 in Figure 5.3,
2. Second between 40 and 50ms, marked as Frame 2-3,
3. And third between 50 and 60ms, Frame 3-4.

As illustrated in Figure 5.3, distribution of the RMS cue does not appear to change as we make measurements in the nasal murmur at different distances from the landmark point. Distributions of the values in three frames have equal standard deviation, with a 0.01dB change in the mean over three frames. In other words, for the nasal murmur at any distance from the landmark point the RMS value between consecutive 10ms time intervals in our implementation of the algorithm is expected to remain below 1 dB[5].



	Frame 1-2	Frame 2-3	Frame 3-4
Mean	0.14	0.14	0.13
St. Dev.	0.12	0.12	0.12

Figure 5.3 – Distribution of the ΔRMS cue when measured at different distances from the nasal sonorant landmark point while in the nasal murmur region.

[5] When examining the change in the RMS value, we remove the unnecessary accuracy used to evaluate the distribution and adhere to Chen's recommendation.

5.5.1.4 Modification of the acoustic criterion for the ΔRMS

In summary, change in the RMS value between two consecutive time frames centered around 30 and 40ms past the estimated landmark point produces statistically different values for nasal and all but the lateral group of non-nasal sonorant landmarks. The means and spread of distributions of three non-nasal groups of sonorant landmarks are well separated from the nasal group. Nasal and lateral groups of sonorant landmarks show a significant overlap in their distribution. For nasal landmarks, the range of values falls fully within the acoustic criterion suggested in Chen [1]. In addition, the cue appears unchanged when measured in nasal murmur, but at different distances from the landmark point. The acoustic criterion that we maintain based on this analysis of the ΔRMS value is:

Portion of the signal for which we hypothesize to exhibit characteristics of the nasal murmur cannot have the RMS value change by more than 1 dB from one 10 ms time interval to the next.

Based on the ANOVA analysis and distribution of values shown in Figure 5.2, the ΔRMS criterion will effectively separate nasal from all but the lateral group of estimated sonorant landmarks in Chen's decision-based algorithm.

Next we analyze the distribution of the f1 values in Frame 2 and the longitudinal characteristics of the cue when measured in nasal murmur at different distances from the estimated nasal sonorant landmark point.

5.5.2 f1

5.5.2.1 Distribution of f1 values across landmark types

Figure 5.4 illustrates the distribution of f1 values for the nasal and each group of non-nasal sonorant landmarks. As a reminder to the reader f1 denotes the largest prominence in the 0-788 Hz frequency range as defined by Chen. f1 values measured with automated algorithms around nasal landmarks confirm the requirement observed in past studies that the energy of the nasal murmur is centered around 250Hz. Visual inspection of Figure 5.4 indicates that the landmarks inserted at semivowel boundaries show the same characteristics as nasals – the distribution of f1 values for this group of sonorant landmarks falls almost completely within the nasal distribution. As with the ΔRMS cue, sonorant landmarks inserted within vowel segments have the widest distribution of f1 values. This characteristic of the vowel distribution for f1 can be attributed to the difference in the vowel type and position of the estimated landmark within the vowel segment. As the value of the true first formant changes, the value of the largest harmonic in the 0-788Hz range, used as f1 in our study, also varies significantly. In addition, when measured at the beginning or end of an utterance, f1 appears to produce significantly lower values, often below 100 Hz.

Using alpha of 0.05 in the ANOVA statistical analysis function indicates that the distribution of f1 values is statistically different for nasal sonorant landmarks and landmarks at lateral and obstruent boundaries. As confirmed during the visual inspection, the mean and spread of the f1 distribution for landmarks at semivowel boundaries are not well separated from the nasal group. Pair-wise statistical analysis between nasal and vowel groups indicates that the means of the two distributions are not well separated despite the large difference in the spread.

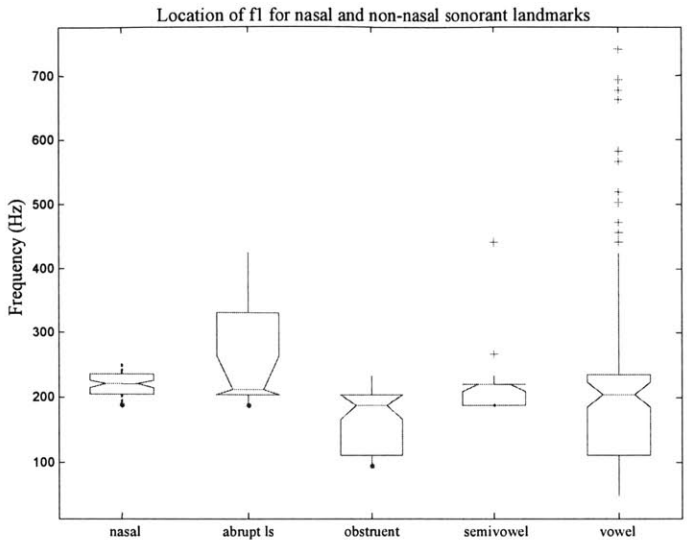


Figure 5.4 – Range of values observed for the f1 cue when measured at 70 nasal and 181 non-nasal sonorant landmarks.

Table 5.6 gives an overview of the statistical analysis results for f1 for nasal and each group of non-nasal sonorant landmarks as classified in Chapter 2 and Appendix A.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Abrupt [l]	70	14	0.001	**Reject H_0**
Nasal	Obstruent	70	50	1.056×10^{-11}	**Reject H_0**
Nasal	Semivowel	70	18	0.409	**Cannot reject H_0**
Nasal	Vowel	70	98	0.489	**Cannot reject H_0**

Table 5.4 – Results of the statistical analysis for f1 for nasal and non-nasal sonorant landmarks, sub-classified across the four groups defined in Table 2.3. The results indicate that measuring f1 at nasal and obstruent landmarks will produce statistically different values. Distribution of values for the lateral landmarks appears also to be significantly different than nasal, though this observation is contingent on further examination of a larger number of tokens.

5.5.2.2 Distribution of f1 values for nasal sonorant landmarks

The range of f1 values for the nasal group of estimated sonorant landmarks falls within the $126 \leq f1 \leq 347$ range specified by Chen. Disparity between the two ranges could

be attributed to the difference in the database structure – acoustic characteristics within a VCV database are considered better behaved, possibly leading to a smaller standard deviation in the observed values for this study than in Chen. Another possibility is that Chen's criteria are not tailored for different speakers and contexts. A minor point that could also contribute to the smaller standard deviation observed in this study is our decision to enhance the frequency resolution to 15.75HZ from the original resolution of 31.5Hz used in Chen. Based on this analysis, the f1 criterion is not responsible for the errors observed in the algorithm's performance.

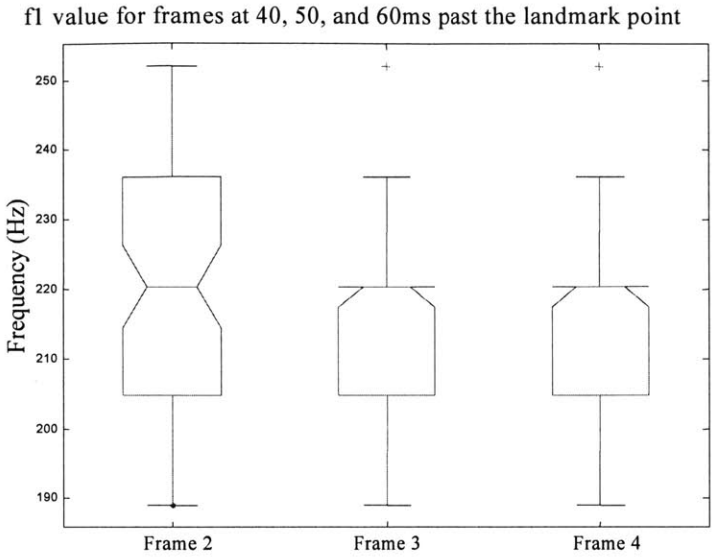
	Nasal	
	Automated	Chen
Min.	189.01	126
Max.	252.02	347
Mean	217.36	--
St.Dev.	17.51	--

Table 5.5 – The range of values and statistical information for the observed parameter f1. The mean value of 217.4Hz confirms the requirement that the energy within the nasal murmur is centered around 200-300Hz.

5.5.2.3 Longitudinal characteristics of the f1 cue

We next examine the longitudinal behavior of this cue using the previous observation that each nasal sonorant landmark is adjacent to the nasal murmur that extends at least 80ms past the landmark point. The three observations of f1 in Figure 5.5 are measured in time frames centered around 40, 50, and 60ms past the nasal sonorant landmark point. From the box plot and tabulated data, it appears that the mean and standard deviation of the f1 value vary as we repeat the measurement at different distances from the nasal sonorant landmark. Noting that the algorithm used to extract the f1 value has a frequency resolution of 15.75Hz (obtained by calculating the 1024-point DFT of the signal sampled at 16129Hz), allows us to conclude that variation in the mean

(<1.5Hz over 30ms) and standard deviation (~2Hz over the same interval) are relatively insignificant with respect to the frequency resolution.



	Frame 2	Frame 3	Frame 4
Mean	217.36	216.46	216.01
St. Dev.	17.51	16.68	15.44

Figure 5.5 – Distribution of the observed f1 values for three time frames, centered around 40, 50, and 60ms respectively. The means and standard deviation appear to be stable as a function of the distance from the estimated landmark point.

5.5.2.4 Modification of the acoustic criterion for f1

Formulating the f1 acoustic cue in terms of an acoustic criterion within the nasal detection module opens the discussion whether the observed f1 should fall within the range specified by Chen or the one observed in this study. A possible approach in formulating this acoustic criterion is to require that the observed f1 value fall within three standard deviations away from the mean of the aggregated distribution of values from three frames. By including three standard deviations in our criterion we guarantee to

include 99% of all samples in a normally distributed population. A signal portion that has an f1 peak falling within $167 \leq f1 \leq 267$Hz in this study is considered to show properties of the nasal murmur. Qualitatively, the difference in the range observed in this study and in Chen implies that this acoustic cue is possibly speaker and context dependent, and that a greater variety of databases will further modify its structure.

The f1 acoustic criterion requires that the calculated f1 value falls in the $165 \leq f1 \leq 267$ Hz range.

From the ANOVA analysis and distributions illustrated in Figure 5.4, it appears that the formulation of an acoustic criterion based on the f1 cue will discriminate to a varying degree between nasal and all but the semivowel group of sonorant landmarks in Chen's decision-tree algorithm. The cue will be somewhat effective at rejecting outliers in the semivowel and vowel groups of estimated sonorant landmarks.

In the next four sections we analyze the spectral tilt characteristics by examining the distribution of A_1-A_2, A_1-A_3, A_2-A_3, and SAD in Frame 2 for nasal and each group of non-nasal sonorant landmarks. In addition, we examine the longitudinal behavior of the cues when measured at different distances from the landmark point.

5.5.3 A_1-A_2

5.5.3.1 Distribution of A_1-A_2 values across landmark groups

Figure 5.6 illustrates the results of measuring the A_1-A_2 acoustic cue in Frame 2, centered around 40ms past the landmark point, for nasal and non-nasal sonorant landmarks. Positive A_1-A_2 values for the nasal group of sonorant landmarks with a median of about 32dB, confirm that the spectral contour of the nasal murmur is tilted toward low frequencies. Visual inspection of the distributions reveals large spread for the landmarks inserted within vowel segments and significant overlap of the A_1-A_2 distributions for the remaining groups of estimated sonorant landmarks. Analyzing the significance of this cue with the ANOVA statistical tool and setting alpha to 0.05 allows us to conclude that the distribution of A_1-A_2 values for nasal sonorant landmarks statistically differs from the distribution of landmarks inserted at obstruent boundaries and within vowel segments.

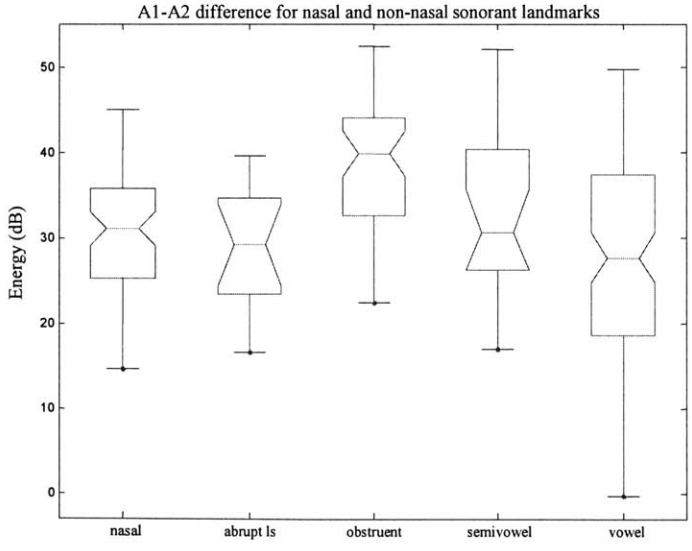


Figure 5.6 – Distribution of the A_1-A_2 values for nasal and non-nasal sonorant landmarks. Pair-wise analysis of the difference between the distribution of A_1-A_2 values indicates that the mean of the nasal sonorant landmark group is significantly separated only from landmarks inserted at obstruent boundaries and within vowel segments.

In the pair-wise analysis between nasal and obstruent groups of sonorant landmarks, the low p-value shows that the means of the two distributions are well separated, despite the large overlap in the observed values. The low p-value for the nasal and vowel analysis is a combination of somewhat similar means and a considerable difference in the distribution spread. High p-values for landmarks at lateral and semivowel boundaries indicate that the two groups have means that are almost identical to the mean of the nasal sonorant landmarks. The range of A_1-A_2 values for the lateral group, moreover, falls completely within the nasal spread. The analysis details are encapsulated in Table 5.6.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Abrupt [l]	70	14	0.244	**Cannot reject H_0**
Nasal	Obstruent	70	50	3.326×10^{-7}	**Reject H_0**
Nasal	Semivowel	70	18	0.26	**Cannot reject H_0**
Nasal	Vowel	70	98	0.034	**Reject H_0**

Table 5.6 – Results of the ANOVA statistical analyses between nasal and each group of non-nasal sonorant landmarks. Setting alpha to 0.05 shows that the A_1-A_2 cue successfully separates nasal sonorant landmarks from those inserted at obstruent boundaries and within vowel segments.

5.5.3.2 Distribution of A_1-A_2 values for the nasal group of sonorant landmarks

The range of values obtained with automated algorithms violates the criterion proposed by Chen: the minimum observed value meets the lower bound of the suggested criterion, while the maximum exceeds it as illustrated in Table 5.7. A range of positive values for this acoustic cue is a consequence of the described spectral tilt – nasal murmur is characterized by a sudden drop in the energy at frequencies above 1000 Hz, which

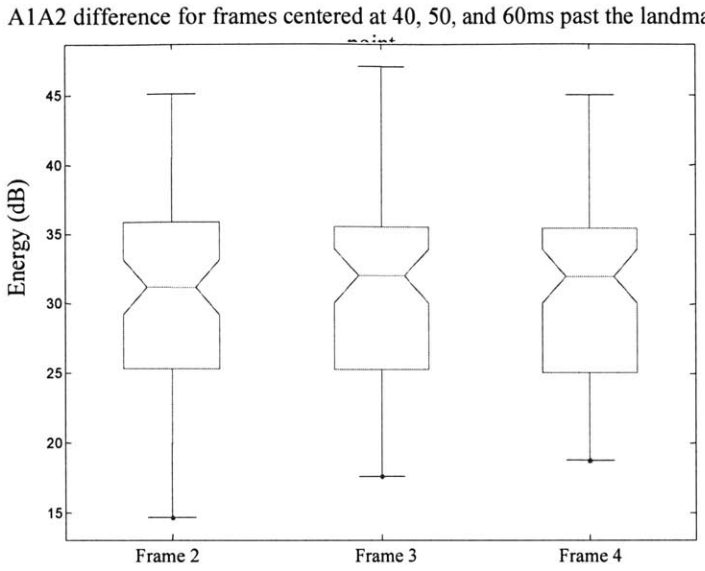
causes the energy of the lower frequency peak A_1 to be consistently greater than A_2. With a disparity between the observed A_1-A_2 values and the suggested criterion, this cue appears to be responsible for some of the errors seen in the performance of the automated implementation of Chen's algorithm for this VCV database.

	Nasal	
	Automated	Chen
Min.	14.71	11.8
Max.	45.04	43.3
Mean	31.06	--
St.Dev.	7.10	--

Table 5.7 – Distribution of values for A_1-A_2 for the nasal sonorant landmarks measured in the frame centered around 40ms past the landmark point. The lower range of values for this acoustic cue is within the limit suggested by Chen, while the maximum exceeds it.

5.5.3.3 Characteristics of A_1-A_2 at different distances from the nasal landmark

We next examine the behavior of this acoustic cue when measured in the nasal murmur, but at different distances from the landmark point. Using the portion of the signal following each nasal sonorant landmark we examine the distribution of the A_1-A_2 cue when measured in the time frames centered around 40, 50, and 60ms respectively, while still in the nasal murmur region. As illustrated in Figure 5.7, distribution of the A_1-A_2 values does not appear to change significantly when measured at different distances from the estimated landmark point while in the nasal murmur region. The mean of the measured A_1-A_2 for a nasal sonorant landmark changes by <0.5dB when measured across three frames. Stability of the cue indicates that once formulated in terms of an acoustic criterion, A_1-A_2 can be used to verify the presence of the nasal murmur at any distance from the estimated landmark point.



	Frame 2	Frame 3	Frame 4
Mean	31.06	31.35	31.52
St. Dev.	7.10	6.59	6.57

Figure 5.7 – Distribution of the A_1-A_2 values measured in three frames from the nasal sonorant landmark in the nasal murmur region. The cue appears to have a relatively stable distribution when measured in the nasal murmur. Change in the standard deviation of the spread is approximately 0.5dB, while the mean varies by <0.5dB over three measured frames.

Agreement with Chen's lower threshold raises the question of how relevant is it that we define a maximum for the A_1-A_2 acoustic cue, as its primary role is to verify that the energy of the peaks in the 788-2000 Hz band is lower than the energy in the 0-788 Hz band. Adhering to the decision-based structure, the approach we take in this study is to establish the minimum by which A1 has to be greater than A2 for a sonorant landmark to pass this criterion for the nasal murmur. Taking a lower limit set by subtracting three standard deviations from the mean puts the two ranges in agreement and formulates the following acoustic criterion:

Portion of the signal for which we hypothesize to exhibit characteristics of the nasal murmur must have a peak in the 0-788 frequency band that is at least 11.1 dB greater than the largest peak in the 788-2000 frequency band.

Because of the large overlap of distributions for nasal and all groups of non-nasal sonorant landmarks, using this acoustic cue in the decision tree algorithm proposed by Chen will separate nasal sonorant landmarks only from those inserted within vowel segments. Arriving at an acoustic criterion that puts this study and Chen in agreement effectively expands the number of examined utterances and makes for a more general criterion. The ANOVA analysis indicates that a statistical approach to formulating an acoustic criterion based on the A_1-A_2 cue may have yielded better results than the decision-tree structure suggested in Chen.

5.5.4 A1-A3

5.5.4.1 Distribution of A_1-A_3 values across landmark types

A_1-A_3 imposes further requirements on the spectral contour of the signal around the landmark point. As illustrated in Figure 5.8 the distribution of A_1-A_3 shows extensive similarities across the five groups, which all span a range of positive values. Visual inspection of the graph indicates that distributions of A_1-A_3 values for three of four non-nasal groups of sonorant landmarks fall fully within the nasal range. The means of distributions, however, appear to be well separated.

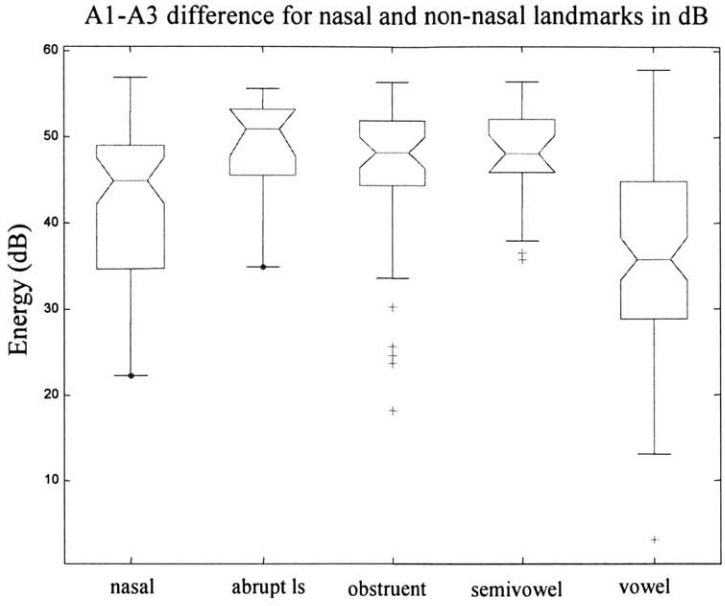


Figure 5.8 – Distribution of the A_1-A_3 values for a set of 251 nasal and non-nasal sonorant landmarks. The box plot visually indicates that the A_1-A_3 means for the groups are well separated, but that distributions for all groups show significant overlap.

The ANOVA statistical analysis, with alpha set to 0.05, confirms the separation of means for the nasal and each group of non-nasal sonorant landmarks in this single frame. Calculated p-values range from 0 for the pair-wise analysis of nasal sonorant landmarks and those inserted within vowel segments, to 0.044 for nasals and landmarks inserted at obstruent boundaries. These low p-values do not reflect the substantial overlap of distributions across the five groups, but mainly indicate that the group means are well separated. Table 5.8 summarizes the details of the statistical analysis for each class of estimated sonorant landmarks.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Abrupt [l]	70	14	0.010	**Reject H_0**
Nasal	Obstruent	70	50	0.044	**Reject H_0**
Nasal	Semivowel	70	18	0.016	**Reject H_0**
Nasal	Vowel	70	98	0.000	**Reject H_0**

Table 5.8 – ANOVA statistical analysis details and results for the A_1-A_3 acoustic cue. A p-value that is smaller than 0.05 across all groups indicates that this acoustic cue is a significant measure when separating nasal from the non-nasal group of sonorant landmarks.

5.5.4.2 Distribution of A_1-A_3 values for the nasal group of sonorant landmarks

Compared to the range of values proposed in Chen, A_1-A_3 values extracted around nasal sonorant landmarks in this study exceed both thresholds, as illustrated in Table 5.9. The mean A_1-A_3 value for the nasal group in Frame 2 is 42.5dB. While the minimum observed value is slightly below the lower permissible bound, discrepancies at the upper threshold are significant. The A_1-A_3 criterion appears to be an important part of the algorithm's erroneous performance.

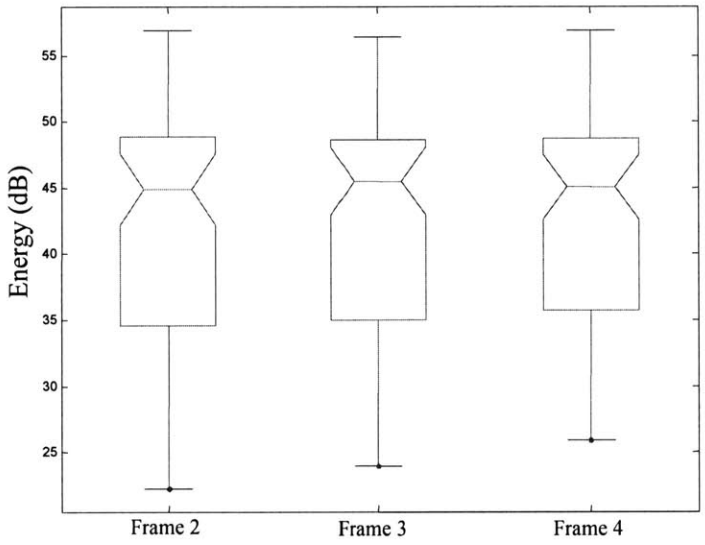
	Nasal	
	Automated	Chen
Min.	22.24	22.8
Max.	56.90	50.3
Mean	42.46	--
St.Dev.	8.65	--

Table 5.9 – Overview of the A_1-A_3 values extracted with automated algorithms against the acoustic criteria for A1-A3 proposed in Chen.

5.5.4.3 Characteristics of A_1-A_3 at different distances from the nasal landmark

Next we employ the same method of examining the longitudinal behavior of the A_1-A_3 acoustic cue when measured at different distances from the estimated nasal sonorant landmark, but within the nasal murmur. Figure 5.9 illustrates the stability of the acoustic cue where the mean and standard deviation remain practically unchanged across the three measured frames. Together with the ΔRMS cue, A_1-A_3 shows greatest stability when measured at different distances from the estimated landmark point. The implication of the longitudinal characteristic is that the A_1-A_3 acoustic cue analysis performed on a single frame is applicable to any nasal sonorant landmark and any frame for this data set.

A_1-A_3 difference for frames centered at 40, 50, and 60ms past the landmark



	Frame 2	Frame 3	Frame 4
Mean	42.46	42.47	42.62
St. Dev.	8.65	8.48	8.61

Figure 5.9 – Distribution of the A_1-A_3 acoustic cue for the three frames centered at 40, 50, and 60ms past the nasal sonorant landmark point. Similarity in the distribution of values for the three frames is indicative that results of the analysis of a single frame can be extended to any nasal sonorant landmark and any frame that shows characteristics of the nasal murmur.

5.5.4.4 Modification of the acoustic criterion for A_1-A_3

Definition of the A_1-A_3 acoustic criterion is similar to characterizing the expectation for A_1-A_2. Because both acoustic cues are projecting the expectation that the selected A_2, A_3 peaks be weaker than A_1, we define A_1-A_3 by setting a minimum the signal needs to exceed to adhere to the nasal murmur characteristics. This approach is also in agreement with the observation that both A_1-A_2 and A_1-A_3 fall within the lower range suggested in Chen, but not the upper. Taking the minimum value to be three standard deviations from the mean defines this acoustic criterion as requiring that A_1-A_3 for nasal sonorant landmarks be at least 16.9 dB. This constitutes the modified acoustic criterion.

The requirement for nasal sonorant landmarks is that A_1 be at least 16.9 dB greater than A_3.

Despite the well-separated means, use of this acoustic cue within a decision-tree algorithm is limited due to the large overlap of the distribution of values for nasal and non-nasal sonorant landmarks. Formulating the acoustic criterion for a decision-tree structure as suggested in Chen, makes this cue effective only when separating nasal sonorant landmarks from those inserted within vowel segments. A statistical approach appears to have been a more effective choice for the A_1-A_3 acoustic cue.

5.5.5 A_2-A_3

5.5.5.1 Distribution of A_2-A_3 values across landmark types

Measuring the A_2-A_3 acoustic cue in the frame centered at 40ms past the landmark point produces almost identical distribution of values across five sonorant landmark

groups as illustrated in Figure 5.10. The range of values for the nasal group is mostly positive, with the median around 13 dB. The means of some non-nasal groups of sonorant landmarks are well separated from the nasal, though their distributions fall almost completely within the nasal range.

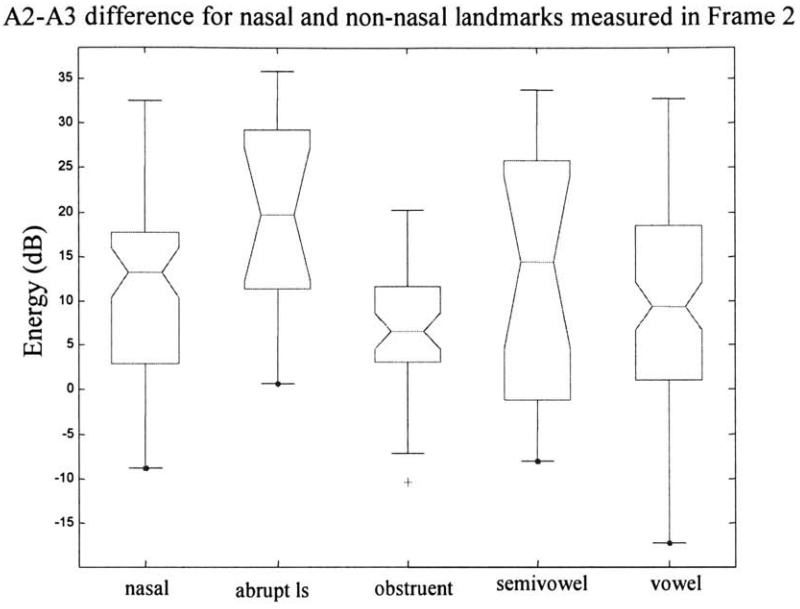


Figure 5.10 – Distribution of A2-A3 values when extracted at nasal and non-nasal sonorant landmarks.

Despite the large overlap, with alpha set to 0.05 the ANOVA function indicates that the distribution of values for A_2-A_3 is statistically different when measured at nasal sonorant landmarks, and at lateral and obstruent boundaries. For the remaining two groups, the means and distributions are not well separated as indicated by the relatively high p-values in Table 5.10.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Abrupt [l]	70	14	0.003	**Reject H_0**
Nasal	Obstruent	70	50	0.020	**Reject H_0**
Nasal	Semivowel	70	18	0.292	**Cannot reject H_0**
Nasal	Vowel	70	98	0.122	**Cannot reject H_0**

Table 5.10 – Summary of the ANOVA statistical analysis details for the acoustic cue A2-A3. The observed mean and distribution of values for the nasal and non-nasal groups of sonorant landmarks are not well separated as indicated by the high p-value.

5.5.5.2 Distribution of A_2-A_3 values for the nasal group of sonorant landmarks

As illustrated in Table 5.11, the measured values for A_2-A_3 for the nasal group of sonorant landmarks fall within the permissible range suggested in Chen. The agreement between the two ranges may suggest that this acoustic cue is somewhat speaker and context independent – examination of the A_2-A_3 values on 70 nasal sonorant landmarks in this study affirms the previously defined range. Chen, however, formulates this acoustic cue based on empirical evidence alone; for this reason, it is possible that the formulation of this acoustic criterion will change with additional databases and contexts.

	Nasal	
	Automated	Chen
Min.	-8.76	-9.7
Max.	32.53	33.6
Mean	11.39	--
St.Dev.	9.63	--

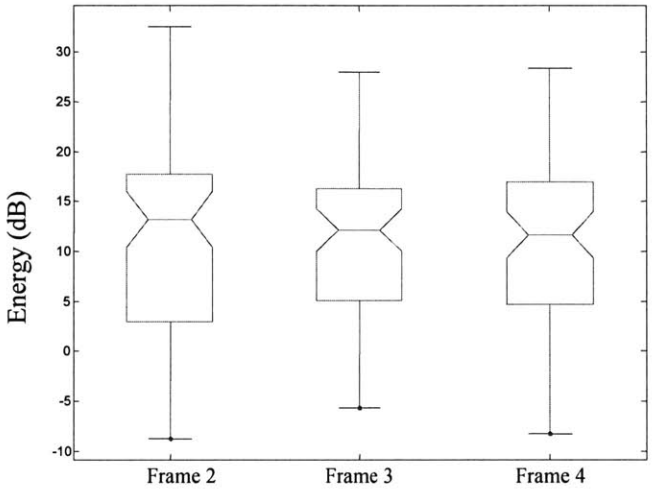
Table 5.11 – Distribution of A_2-A_3 values extracted at nasal and non-nasal sonorant landmarks. The range observed in this study is in agreement with Chen.

5.5.5.3 Characteristics of A_2-A_3 at different distances from the nasal landmark

In this section, we examine the stability of the A_2-A_3 cue when measured at different distances from the landmark point. From Figure 5.11 it appears that the mean

for the distribution of A_2-A_3 values shows no significant variability when measured at different distances from the estimated landmark point. The mean of the three frames varies by 0.25dB over three frames. The spread of the distribution varies to a greater extent than the mean. In this observation the largest spread still meets the suggested criterion, indicating that errors in the performance of the algorithm are not related to the A_2-A_3 criterion. Variability in the standard deviation of A_2-A_3 for different frames within the nasal murmur indicates that more databases and contexts will likely further augment this acoustic criterion.

A_2-A_3 difference for frames centered at 40, 50, and 60ms past the landmark point



	Frame 2	Frame 3	Frame 4
Mean	11.40	11.13	11.15
St. Dev.	9.63	8.85	9.20

Figure 5.11 – Distribution of A_2-A_3 values for the three frames measured by placing the center of the Hamming window at 40, 50, and 60ms respectively. Distribution of values shows some variability in the standard deviation as a function of distance from the estimated landmark point.

5.5.5.4 Modification of the acoustic criterion for A_2-A_3

Chen particularly isolates the A_2-A_3 acoustic cue as being successful at rejecting sonorant landmarks inserted within [i] segments, without specifying its overall contribution to the non-nasal rejection rate. Further examination of A_2-A_3 values for the sonorant landmarks inserted within vowel segments indicates that landmarks within the vowel [i] tend to produce lower (more negative) A_2-A_3 values than the remaining vowels and other non-nasal groups of sonorant landmarks. Without a theoretical basis, we propose to maintain the original criterion for A_2-A_3 defined by Chen as long as it does not contribute to the erroneous performance of the algorithm. As a cue that is based on empirical evidence alone A_2-A_3 will most likely have to be further refined when examined in a greater variety of databases and contexts.

The requirement for nasal sonorant landmarks is that A_2-A_3 falls in the [-9.7db 33.6dB] range.

Comparing the formulated criterion against Figure 5.10 shows that this acoustic cue will reject a small number of landmarks inserted within vowel segments and at lateral boundaries. Based on the results of the ANOVA analysis, it is possible that taking a statistical approach to formulating this criterion would have been more effective.

5.5.6 SAD

5.5.6.1 Distribution of SAD values across landmark types

The last spectral tilt parameter we analyze is the Sum of Amplitude Differences (SAD). As a reminder, Chen defines SAD as

$$SAD = (A_1 - A_2) + (A_1 - A_3) + (A_1 - A_4) + (A_1 - A_5). \quad (5.4)$$

Figure 5.12 shows the range of SAD values measured for the five groups of estimated sonorant landmarks. Visual inspection of the graph indicates a large overlap of the distributions across all landmark groups. Distributions of values for the lateral, obstruent, and semivowel group of sonorant landmarks, furthermore, are fully contained within the nasal distribution, with almost identical medians. The group of sonorant landmarks inserted within vowel segments shows the largest spread, with the range spanning approximately 120dB, compared to 90dB for the nasal and about 40-60dB for the remaining groups.

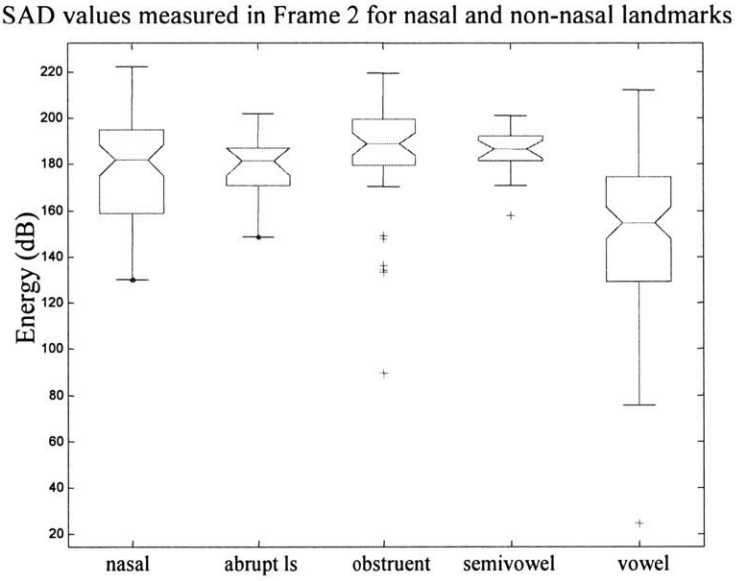


Figure 5.12 – Distribution of SAD values for nasal and all groups of non-nasal sonorant landmarks.

As with previous cues, we next analyze the difference between the five distributions with the ANOVA statistical analysis function. With alpha set to 0.05, the calculated p-values confirm the similarity between the distributions for the nasal and all but the vowel group of non-nasal sonorant landmarks. A very low p-value for the pair-wise analysis of the nasals and vowels indicates a significant difference in the mean and

spread of SAD values for these two groups. Details of the statistical analysis are illustrated in Table 5.12.

Group 1	Group 2	# of tokens in Group 1	# of tokens in Group 2	p-value	H_0: Means from Group 1 and Group 2 are the same
Nasal	Abrupt [l]	70	14	0.837	**Cannot reject H_0**
Nasal	Obstruent	70	50	0.362	**Cannot reject H_0**
Nasal	Semivowel	70	18	0.294	**Cannot reject H_0**
Nasal	Vowel	70	98	9.781×10^{-9}	**Reject H_0**

Table 5.12 – ANOVA statistical analysis of the SAD values indicates that the distribution of the nasal group of sonorant landmarks statistically differs only from the distribution of the vowel group. Visual inspection of the remaining groups confirms a large overlap between the non-nasal and nasal distributions.

5.5.6.2 Distribution of SAD values for the nasal group of sonorant landmarks

While agreeing at the lower threshold, the range of SAD values for the nasal group of sonorant landmarks exceeds the upper bound of the permissible range formulated in Chen. This finding is consistent with the observation that both A_1-A_2 and A_1-A_3, which are included in the calculation of the SAD, exceed their respective criterion. Distribution of SAD values for this frame, as shown in Table 5.13, indicates that the maximum value surpasses the suggested criterion by some 21.5dB or close to one standard deviation of the total spread. Current formulation of the SAD criterion, therefore, is a significant factor in the erroneous rejection of nasal sonorant landmarks for the nasal murmur characteristics.

	Nasal	
	Automated	Chen
Min.	130.25	119.5
Max.	222.30	200.7
Mean	179.83	--
St.Dev.	21.90	--

Table 5.13 – Distribution of values calculated for SAD for the nasal and non-nasal group of sonorant landmarks. While the lower range of values extracted in this study falls within the range suggested by Chen, the upper range exceeds the allowed maximum.

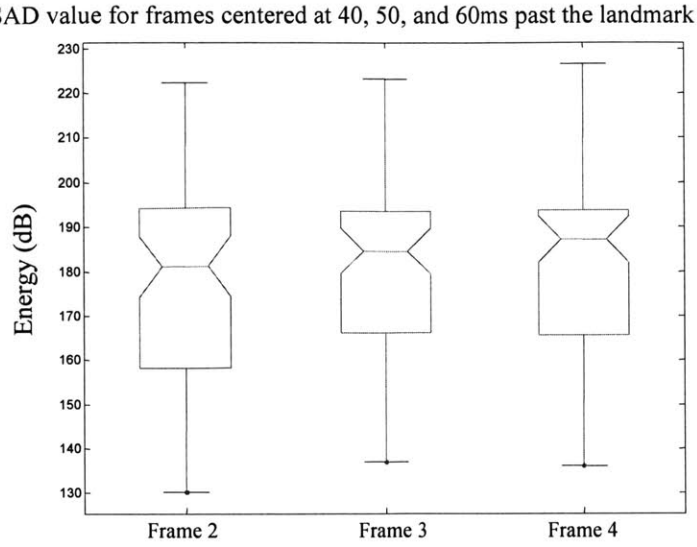
5.5.6.3 Characteristics of SAD at different distances from the nasal landmark

Lastly we characterize the change in the SAD values when measured at different distances from the nasal sonorant landmark, while still in the nasal murmur region. In Figure 5.13 we show the distribution of SAD values for three frames centered at 40, 50, and 60ms past the nasal sonorant landmark point for 70 tokens. Slight variability of the distribution mean across the three frames can be attributed to the individual variability of the five cues that constitute the SAD value. Compared to the 20-21dB standard deviation, a 2dB change in the distribution means over three frames is proportional to the variability of the remaining spectral tilt cues, such as A_1-A_2 and A_2-A_3.

5.5.6.4 Modification of the acoustic criterion for SAD

At this point we have already formulated the acoustic criteria for three of five acoustic cues that make up the SAD value. Two of these criteria have been defined in terms of the minimum value that a sonorant landmark needs to meet in order to be considered as showing nasal murmur characteristics. The same reasoning that we followed in formulating the A_1-A_2 and A_1-A_3 criteria applies to the remaining cues, A_1-A_4 and A_1-A_5, and consequently to the SAD criterion.

Definition of these four cues in terms of a minimum a sonorant landmark needs to meet to adhere to the nasal murmur characteristics verifies that the spectral contour of the signal is skewed towards low frequencies in the same manner as the spectra of the nasal murmur.



	Frame 2	Frame 3	Frame 4
Mean	179.53	180.48	181.34
St. Dev.	21.90	20.92	21.10

Figure 5.13 – Distribution of SAD values for the three frames measured by placing the center of the Hamming window at 40, 50, and 60ms past the landmark point respectively. The spread of values shows some variability in the distribution mean as a function of distance from the estimated landmark point.

Selecting the minimum to be three standard deviations below the calculated mean, the acoustic criterion for SAD requires that sonorant landmarks have the SAD of at least 116.7 dB.

SAD criterion requires that a sonorant landmark have the SAD value of at least 116.7 dB for nasality considerations.

Addition of the SAD cue to the nasal detection module completes the set of requirements imposed on the spectral tilt of the signal around the landmark. Based on the ANOVA statistical analysis and Figure 5.12, the SAD criterion will separate nasal from landmarks inserted within vowel segments only.

5.6 Effectiveness of the modified criteria

In section 5.4 we analyzed the values and effectiveness of six acoustic cues from Chen's algorithm, and suggested modifications that would improve its performance. The proposed modifications attempted to reconcile the observations from the current study and Chen, in an attempt to define a set of acoustic criteria that would be applicable to both examined databases and a greater variety of speakers and contexts. Table 5.14 combines the results of the ANOVA analysis and measured values to illustrate the effectiveness of the modified acoustic criteria in rejecting specific non-nasal groups of estimated sonorant landmarks.

	Vowel	Semivowel	Lateral	Obstruent
ΔRMS	x	x		x
f1	x		x	x
A1-A2	x			
A1-A3	x			
A2-A3	x		x	
SAD	x			

Table 5.14 – Summary of the effectiveness of each cue when rejecting the four non-nasal groups of estimated sonorant landmarks. An 'x' in the f1 entry for 'Vowel' means that the current definition of the f1 criterion will reject some sonorant landmarks inserted within vowel segments. We do not quantify the effectiveness of the criteria.

For some acoustic cues we noted that the statistical approach might have been a better solution when formulating acoustic criteria based on the promising results of the ANOVA statistical analysis. A statistical classification of sonorant landmarks in the nasal or each non-nasal group based on the six selected cues would not depend on firm thresholds; based on the selected method, rather, the classification would give a confidence rating with which a landmark can be classified in any one group of estimated sonorant landmarks from Table 2.3.

5.7 Performance of Chen's algorithm with the modified acoustic criteria

Using the original Chen algorithm structure, but this time with the modified acoustic criteria as explained in Section 5.4 eliminates the instances where nasal sonorant landmarks are rejected as non-nasal. This improvement in the performance, however, is accomplished at the expense of the non-nasal rejection rate – with the modified acoustic criteria now, 70.6% compared to the original 86.1% of non-nasal sonorant landmarks are rejected for the nasal murmur characteristic with the same algorithm. The largest decrease in the performance is seen in the rate at which sonorant landmarks at semivowel and lateral boundaries are rejected by the algorithm. This decrease appears to be mainly a consequence of eliminating the upper limit on the A_1-A_3 value range in the modified acoustic criteria. Table 5.15 summarizes the number of tokens in each class that was tested with the modified acoustic criteria and the rejection rate for each class of estimated sonorant landmarks.

Class of Sonorant Landmark	Comment	Total [s] landmarks	Landmarks rejected for nasality	Rejection Rate
Nasal	True positive	70	0	0.00%
Abrupt l		14	5	35.71%
Vowel	False positive	98	83	84.69%
Semivowel		18	5	27.78%
Obstruent		50	34	68.00%
Total	**Nasal**	70	0	0.00%
	Non-nasal	180	127	70.56%

Table 5.15 – Performance analysis of Chen's original algorithm implemented with the modified acoustic criteria suggested in Section 5.4. The modified criteria show in instances where nasal sonorant landmarks are rejected for the nasal murmur characteristics. This improvement in the performance, however, is accomplished at the expense of the non-nasal rejection rate – with the modified acoustic criteria, 70.6% compared to the original 85.6% of non-nasal sonorant landmarks are rejected for the nasal murmur characteristic with the same algorithm.

5.8 Summary

In this chapter we discuss and formulate an acoustic criterion for the nasal murmur by examining and modifying Chen's algorithm for the nasal murmur detection. The algorithm examines six acoustic cues in the region of the signal adjacent to the sonorant landmark with a lower energy level and determines whether they adhere to the nasal murmur expectation formulated in terms of the acoustic criteria.

Performance of the reconstructed algorithm is characterized by a high rejection rate across all five groups of estimated sonorant landmarks. For the non-nasal groups, the algorithm rejects 85.6% landmarks as not showing characteristics of the nasal murmur. The efficiency of the algorithm is offset by a high error rate; with the original criteria Chen's algorithm also rejects 34.3% of nasal sonorant landmarks as not showing characteristics of the nasal murmur.

Analysis of the cue values in a frame centered at 40ms past the estimated sonorant landmark point allows us to identify the acoustic criteria that cause errors in the algorithm performance and the ways in which they can be corrected. Analyzing the same set of sonorant landmarks with the original algorithm structure and the modified criteria results in no nasals being rejected for the nasal murmur characteristics. For non-nasal groups of sonorant landmarks the algorithm rejects 70.6% landmarks as not showing characteristics of the nasal murmur.

Chapter 6

6.1 Nasalized Vowel

As previous studies indicate, nasalization of the vowel preceding a nasal segment provides valuable cues to the presence and detection of the [nasal] feature. In a 1988 study, Stevens, Andrade, and Viana cite the requirement for a various degree of vowel nasalization in judging the naturalness of the adjacent nasal segment across French, Portuguese, and English languages [21]. The most natural sounding synthesized utterances for native speakers of American English included both elements of nasal murmur and vowel nasalization, thus giving direct evidence that the two acoustic manifestations often appear concurrently in the nasal segment production and recognition. This chapter describes the acoustic cues found characteristic of the vowel nasalization in the past studies and formulates them in such a way that they can be extracted with automated algorithms. The cues are mainly based on Chen and examine the portion of the signal with a higher energy level around the sonorant landmark for signs of spectral prominences, termed P_1 and P_0, due to the nasal cavity. In this chapter we:

1. Describe the suggested acoustic cues,
2. Implement them in terms of automated algorithms in MATLAB and evaluate on 250 nasal and non-nasal sonorant landmarks from Chapter 2,
3. Attempt to evaluate their effectiveness in separating nasal from non-nasal landmarks and propose an acoustic criterion that can be used within the nasal detection module.

6.2 Acoustic cues for nasalized vowels

American English makes no phonemic difference between nasalized and oral vowels, thus allowing speakers to nasalize vowels in any phonetic context and to a varying extent. In this study we use the term 'nasalized vowel' to refer to those vowels that are adjacent to a nasal segment, either preceding a nasal closure or following a nasal release. In choosing this notation we assume that vowels adjacent to nasal segments will be nasalized to a greater extent and with a greater consistency among speakers. Non-nasalized vowels, by the same token, are all other vowels in the VCV database. In formulating an acoustic criterion for nasalized vowels, we use Chen's proposed algorithm because it appears to encapsulate a number of previous studies and observations when suggesting two characteristic acoustic cues that can be adjusted for the vowel type [1], [2], [3]. In this section we describe the details of the proposed cues and algorithm.

Amplitude of the first formant, A_1, and prominences due to the nasal cavity, P_0 and P_1

Theoretical basis	Chen claims that three parameters capture the presence of extra peaks in nasalized vowels: 1. An extra peak between the first two formants with amplitude P_1, 2. One at lower frequencies, often below the first formant, with amplitude P_0, 3. Amplitude of the first formant A_1 (estimated by the energy of the peak closest to the calculated first formant). Theoretically, during the production of a nasalized vowel, the amplitude of the first formant is expected to weaken compared to the equivalent oral vowel, with A_1 lowering by around 5 dB. In contrast, the extra peaks due to the coupling to the nasal tract, P_1, and the sinuses, P_0, can increase for nasalized vowels by around 13 dB and 3 dB respectively. Theoretically then, adjusted parameters A_1-P_1* can differ by as much as 18 dB and A_1-P_0* by 8dB between

	nasal and oral vowels.
Quantitative form	Chen proposes the following algorithm for the measurement of P_0 and P_1: - Estimate the amplitude of P_0 by measuring the energy of the larger of the first two harmonics, - Observe P_1 by measuring the energy of the largest of three peaks around 1 kHz. For A_1, Chen hand-estimates the frequency of the first formant and approximates A_1 by measuring the energy of the peak closest to the calculated formant. Chen measures the mean difference of $D_1=A_1-P_1$* for oral and nasal vowels is 10-15 dB, while the $D_0=A_1-P_0$* difference has a range of 6-8 dB.
Algorithm	Designing an algorithm for the successful characterization and extraction of the vowel nasalization parameters involves a difficult task of designing a fully functional, automated formant tracker, which would be a thesis project in its own right. Instead of using a fully automated formant tracker then, a semi-automated model is used which allows the calculated values for F1 through F3 to be hand-corrected. The semi-automated model uses the COLEA formant tracker to calculate the formant values based on the LPC analysis of the underlying speech segment. Because the LPC analysis of speech signal is limited and often inaccurate, addition of the correction scheme developed in the Speech Communication group at MIT allows formants to be hand-augmented after viewing the corresponding broadband spectrogram of the signal. The adjustment formula proposed in Chen also involves the estimation of the formant bandwidth, which cannot be measured with the semi-automated model [3]. Here we use the bandwidth as calculated by the COLEA software tool, but bear in mind that the adjustment formula could be inaccurate due to this measure. P_0 is calculated by selecting the larger of the first two peaks located with the peak-picking function described in previous sections. P_1 is estimated through the steps suggested by Chen; we first determine the energy of the three peaks around 1000 Hz and select the peak with the largest energy to be P_1. A_1 and F1 are calculated by taking the energy of the largest peak next to the estimated first formant frequency, while F2 is estimated by evaluating the energy of the largest peak next to the second formant frequency estimated with the semi-automated formant tracker. The final parameters D_1 and D_0 are calculated using the following formulae: $D_1 = A_1 - P_1 - T1_{approx.} - T2_{approx.}$, (6.1)

	and $D_0 = A_1 - P_0 - T1(F_{P_0}) - T2(F_{P_0}), (6.2)$ where $T1_{approx.} = \frac{F_1^2}{(F_{P1} - F_1) \times (F_{P1} + F_1)}, (6.3)$ $T2_{approx.} = \frac{F_2^2}{(F_2 - F_{P1}) \times (F_{P1} + F_2)}, (6.4)$ $T1(F_{P_0}) = \frac{(0.5B_1)^2 + F_1^2}{\sqrt{[((0.5B_1)^2 + (F_1 - F_{P_0})^2) \times ((0.5B_1)^2 + (F_1 + F_{P_0})^2)]}}, (6.5)$ $T2(F_{P_0}) = \frac{(0.5B_2)^2 + F_2^2}{\sqrt{[((0.5B_2)^2 + (F_2 - F_{P_0})^2) \times ((0.5B_2)^2 + (F_2 + F_{P_0})^2)]}}. (6.6)$ Adjustment formulae are valid under the assumption that the bandwidths of the formants are much less than the formant frequencies. The three parameters are extracted in four time frames, with the first window centered at 60ms before a sonorant closure and 60ms past a sonorant release, and with each subsequent window 10ms closer to the landmark point. Because D_1 and D_0 measure a difference between spectral prominences, they are independent of the speech intensity for a specific utterance.

6.3 Analysis of the measured data

The range of values extracted for the D_0 and D_1 acoustic cues did not show the same values as observed in Chen [3]. Values appeared to be inconsistent with the theoretical basis and between multiple trials on the same utterance. The factors that most likely caused inconsistent observations are bandwidth estimations from COLEA that relied on the LPC coefficients, irregular peaks found in the region of the first and second formants, and lack of accuracy of the semi-automated formant tracker. Though we

include bandwidth error as a possible cause in the observed D_0 and D_1 values, it is fairly difficult to evaluate its contribution to the overall calculation of the two parameters. More easily observed inconsistency was found in the location and energy of the three estimated peaks around 1000Hz. For some utterances the calculated DFT magnitude showed irregular peaks that often appeared exactly between harmonics and that greatly skewed our estimation of the location of the spectral prominences. This estimation inaccuracy was worsened with irregular peaks common in unsmoothed short-time DFT spectra in the second formant region. Estimation of F_1 and F_2, however, proved to be the hardest task in the D_0 and D_1 calculation. In the next section we describe the functionality of the semi-automated formant tracker and the reasoning behind our recommendation that vowel nasalization cues be excluded from the nasal detection module.

6.4 Semi-automated formant tracker functionality

The Semi-automated Formant Tracker developed in the Speech Communications group at MIT is a tool that allows users to modify the results of the LPC analysis to better align with signal's true formant values. Selection of this formant tracker was based on the observation that available software tools, such as COLEA or Praat, yielded inaccurate results when faced with nasal segments and nasalized vowels, due to the additional nasal poles. The idea of the semi-automated formant tracker is rather attractive. At the input, the tracker takes a file containing the LPC coefficients for the signal calculated either by the Praat or COLEA Formant Tracker to produce the following display:

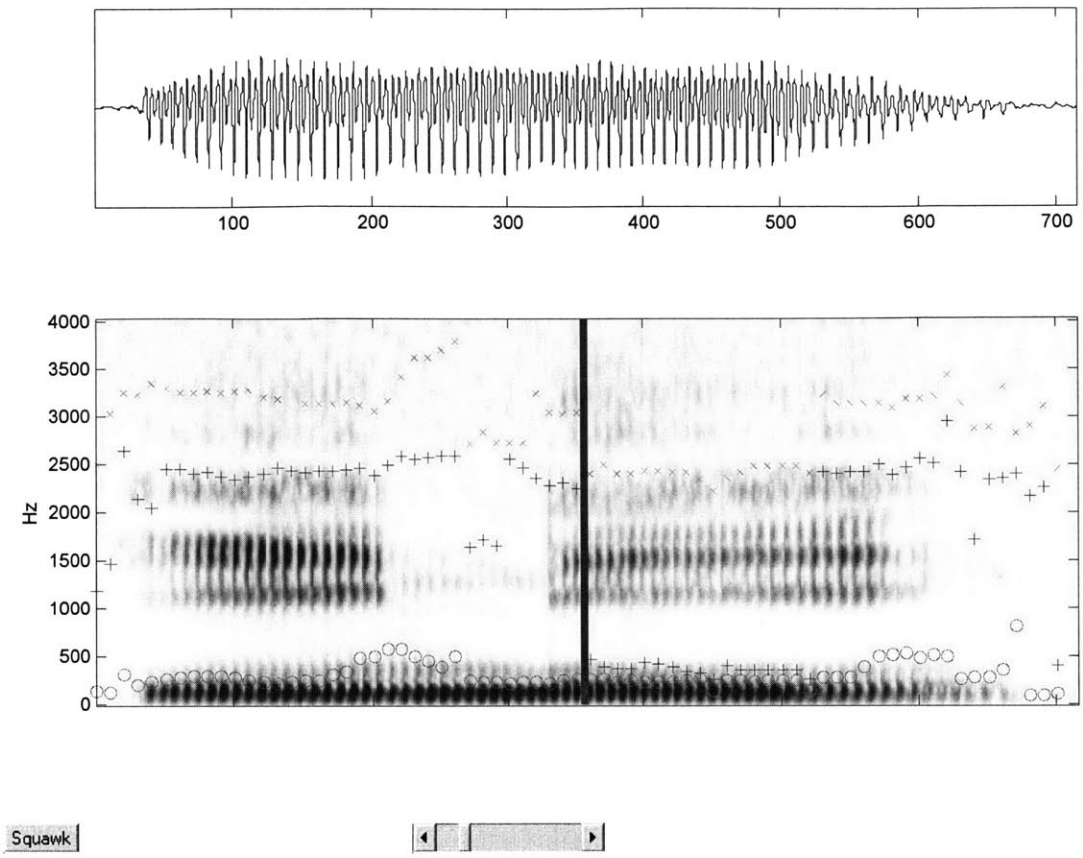


Figure 6.1 – Graphic interface for the semi-automated formant tracker developed at MIT's Speech Communications Laboratory. The two graphs are signal waveform, at the top panel, and broad spectrogram with the LPC coefficients calculated by the COLEA Formant Tracker.

The two graphs in Figure 6.1 are illustrations of the signal waveform and broadband spectrogram for the utterance [ini]. Overlaid on the spectrogram are the 16^{th} order LPC coefficients produced by tracking the formants in COLEA. The displayed LPC coefficients indicate frequencies of the first three formants; 'o' represents the first estimated formant in the LPC analysis, '+' the second, and 'x' the third. Moving the slider to any LPC coefficient or any time frame allows the user to shift its value to what he or she believes represents the true formant. The tracker allows for multiple correction of the same coefficient and does not force the corrected value to fall on a calculated DFT magnitude peak or harmonic. In some instances, correcting the LPC coefficients based on

the broadband spectrogram alone allows experienced users to estimate the corrected values relatively accurately. For other utterances, for example the utterance [ini] shown in Figure 6.1, the user is faced with two options:

1. The first is to estimate the value to the best of one's ability,
2. The second option is to display the short-time spectra with some other software tool in order to correctly determine the exact location of the estimated formant value.

Using the first approach for a single user showed significant variation in the labeling of the same utterance in multiple trials. Variation in the F2 values was between 50 to 200 Hz, in many instances estimating the second formant frequency to be one or two harmonics different than in a previous trial. Estimation was further complicated by the presence of the subglottal resonance around 1500 Hz. Estimating the value of the first formant based on the spectrogram alone also proved to be challenging, as it was difficult to separate the formant from the subglottal resonance at lower frequencies. It is predicted that the inconsistency in the F_1 and F_2 measurements would only increase when faced with multiple users with a varying degree of experience and accuracy in formant estimation.

Requiring a user to revert to the short-time spectra to correctly determine the correct position of each formant equates this approach to hand-measuring the formant values. A number of past studies that served as the basis for proposing this acoustic cue have already hand-measured and examined the spectral prominences due to the nasal cavity. Addition of this analysis to the automated nasal detection module would significantly

increase its complexity without adding new information to the speech processing community. It is also questionable as to how the variability would change between users who estimated the formants based on the spectrogram alone, versus those who referred to the short-time spectra.

Most importantly, however, the goal of the automated nasal detection module is to provide algorithms that will function with little or no input from the user, in an attempt to guarantee consistency in the way measurements are conducted across trials, contexts, and users. Adding an algorithm that will vary for a single user based on the accuracy level of each trial, and for multiple users based on their experience and consistency level, will defeat the original goal set out at the beginning of this study. We therefore suggest postponing the addition of the vowel nasalization analysis to the nasal detection module until a fully automated formant tracker is available that will provide the required consistency and accuracy in the estimation of F_1, B_1, F_2, and B_2. The currently available software tools cannot be used to track changes in the energy of a specific harmonic or formant for short time intervals.

Addition of this module will also be significantly simplified when the LAFF system starts accessing the lexicon to stipulate possible cohorts of words. With possible word cohorts, the vowel nasalization module will be able to compare the "expected" formant tracks against observations made in the signal to estimate the accuracy to the closest matched formant track.

6.5 Summary

The focus of this chapter was to examine ways in which nasalized vowels can be detected in support of the nasal detection. In the survey of available software tools for tracking formant changes across short time intervals, we found no method that guaranteed the desire accuracy and consistency. Our recommendation thus was to postpone the addition of the vowel nasalization criteria to the nasal module until a fully automated formant tracker is available that will guarantee the required consistency and accuracy across contexts and users.

Chapter 7

7.1 Formulating the Nasal Detection Module

This chapter attempts to encapsulate the results of the analyses from previous chapters by proposing a final design for the nasal detection module and evaluating its performance. The nasal module currently supports effective acoustic criteria for the nasal murmur and nasal boundary that use fully automated algorithms in MATLAB to extract the selected acoustic cues. The criteria within the module can be used separately or in combination depending on the user requirements. The proposed nasal module maintains a file for each criterion that keeps a record of all measurements obtained for the specific acoustic cues.

7.2 Pivots as indicators of the change in signal energy

In Chapter 3 we discussed ways in which sonorant landmarks convey the expectation regarding what portion of the signal should adhere to the specific acoustic criterion. As a reminder to the reader, the sign associated with the sonorant landmark in the CLD indicated the trend in the change of the spectral energy - a [-] sonorant landmark indicated decreasing energy due to a more constricted vocal tract, while a [+] landmark implied increasing energy most often associated with a constriction release. In other words, each sonorant landmark point separated the surrounding signal into a region with a higher and lower energy level. In Chapters 4-6 we used the relative energy level of the signal around the sonorant landmark point to design acoustic criteria for nasality. We required that the portion of the signal with a lower energy level conformed to the acoustic criteria for the nasal murmur, while the higher energy level portion was expected to show

characteristics of the nasalized vowel. The landmark point itself was required to agree with the criteria for the vowel-nasal boundary. Chapter 2, however, suggested that limiting nasal detection to estimated sonorant landmarks would only include 56% of all sonorant closures and releases due to a high miss rate of the sonorant estimation in the Consonant Landmark Detector (CLD). The chapter concluded with the recommendation that the nasal detection include all estimated pivots when deciding on the nasality feature. As a first step in combining results from Chapters 2-6 and applying the formulated acoustic criteria on 1483 pivots estimated for this VCV database, we explain how pivots convey the information regarding what portions of the surrounding signal have the higher and lower energy levels discussed in Chapter 3. In other words we explain what information we can use to determine whether each pivot should be tested as a possible nasal closure or release.

Much like the sign for a sonorant landmark, each pivot in Liu's CLD is associated with another measure of the signal energy termed High-Pass Rate-of-Rise (HPROR) value that can be positive, negative, or a zero. The sign of the HPROR value for a pivot indicates the trend in the change of energy in the 1000-8000Hz frequency range, while the quantitative value approximates the relative magnitude of that change. A negative change in this energy would likely be associated with a closure, while the positive change likely denotes the opening of the vocal tract. It is our interpretation that the zero HPROR value indicates a relatively, not absolutely, constant signal around the pivot time. Based on the information conveyed by the HPROR value, it is possible to establish equivalence between a pivot with a negative HPROR and a [-s] landmark, and a pivot with a positive

HPROR and a [+s] landmark. Sonorant landmark equivalence, however, does not apply to the third pivot type with an HPROR value of zero. In this section we attempt to answer two questions:

1. Should pivots with a zero HPROR value be included in the nasal detection and if so, how do we anticipate characteristics of the signal surrounding this pivot type, as there is no clear indication whether the pivot is located at the time of a sonorant closure or release?
2. Are there cases when the HPROR sign of the pivot does not agree with the assumed configuration of the vocal tract?

Analysis of 142 sonorant closures and releases in the VCV database confirms that most sonorant segments have pivots with HPROR signs that correctly specify whether they are placed at the time of a sonorant closure or release. Figure 7.1 illustrates the distribution of HPROR values for the 137 estimated pivots, hand-classified in two groups based on whether the pivot aligns with a sonorant closure or release.

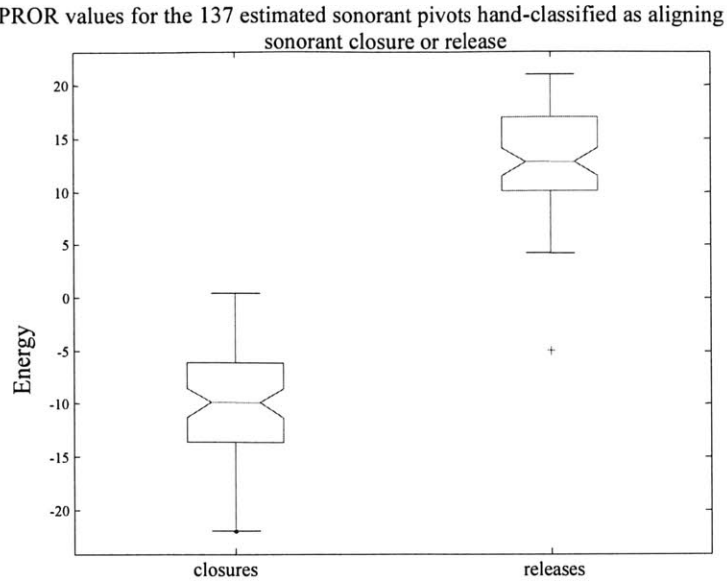


Figure 7.1 – Distribution of HPROR values for 137 estimated pivots hand-classified as aligning with a sonorant closure or release. The values illustrate the general equivalence between the HPROR sign and change in the energy of the signal at the pivot point.

A small number of pivots in both groups appears to violate the equivalence between the HPROR and expected sonorant landmark sign. Further analysis of these exceptions reveals that:

- No pivot aligned with a sonorant closure or release has a zero HPROR value,
- Pivots that violate the rule are located at sonorant closures and releases adjacent to the vowel [u],
- One pivot has a negative HPROR value while aligned with the sonorant release in utterance [uŋu] – this value is shown as an outlier in Figure 7.1,
- Two pivots with positive HPROR values at sonorant closures have values that are close to zero.

The first observation that all exceptions to the rule occur at sonorant closures and releases adjacent to the vowel [u] agrees with our previous hypothesis that the transition

between nonlow, back vowels and sonorant segments is practically non-abrupt due to the vowel's low energy at frequencies above 1000Hz. Consequently, HPROR value of such pivots will be very low and will vary significantly depending on the speaker and context, not always accurately aligning the change in the signal energy with the change in the configuration of the vocal tract. The violation of the HPROR value rule observed in the utterance [uŋu] carries limited significance because the utterance requires difficult positioning of the vocal apparatus that does not exist in American English. Consequently, we do not anticipate occurrence of this type of exception in spontaneous speech or less restricted environments. Our final observation indicates that we can place little confidence in the information relayed by pivots with HPROR values close to zero.

Taking the approach of minimizing the number of missed nasal boundaries leads us to the following method of examining pivots for the nasal characteristics:

1. Pivots with HPROR > 1 are examined for nasal characteristics in the same manner as [+s] landmarks, that is as potential sonorant releases,
2. Pivots with HPROR < -1 are examined as [-s] landmarks in the nasal detection module,
3. Pivots with $-1 \leq$ HPROR ≤ 1 can be aligned with either a sonorant closure or release – as such they have to be examined both as a potential nasal closure and release in nasal detection.

Figure 7.2 illustrates the proposed strategy of analyzing pivots for possible nasal characteristics. From the diagram it is evident that depending on the HPROR value,

pivots can be analyzed as possible nasal closures, releases, or both. An alternative approach is to exclude all pivots with |HPROR| ≤ 1 from further analysis in the nasal detection module. While minimizing the overall number of pivots examined by the nasal detection module, this approach also excludes two pivots at true nasal boundaries from further acoustic analysis in this database.

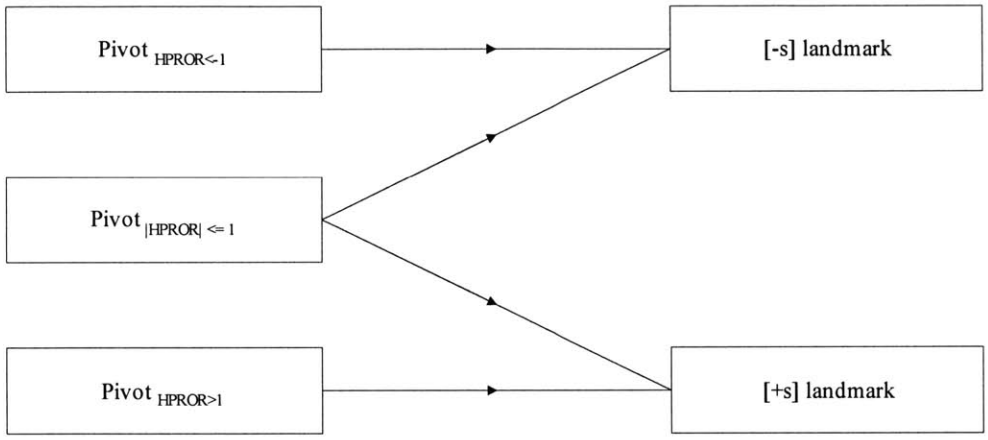


Figure 7.1 - The three pivot types can be equated to the sonorant landmark types based on the HPROR valu pivot with HPROR<1 can be processed as a [-s] landmark as it marks the time of a decreasing energy in the sign A pivot with HPROR>1 can be processed as a [+s] landmark because the energy change associated with this pivo type is consistent with a sonorant release. The minimal energy change associated with pivots with an HPROR value close to zero requires that these pivots be tested as both, potential sonorant closures and releases.

7.3 Pivot contexts

Pivots appear in a greater variety of contexts than sonorant landmarks. As a reminder to the reader, in Chapter 2 we claimed that based on their location within the utterance, all sonorant landmarks can be exhaustively classified into:

1. Landmarks at vowel-nasal/nasal-vowel (VN/NV) boundaries,
2. Landmarks at vowel-lateral/lateral-vowel (VL/LV) boundaries,
3. Landmarks inserted at vowel-obstruent/obstruent-vowel (VO/OV) boundaries,

4. Landmarks inserted at vowel-semivowel/semivowel-vowel (VS/SV) boundaries,
5. Landmarks inserted within vowel segments, and
6. Landmarks inserted in semivowel segments.

Apart from these, pivots can also exist:

7. At vowel-h/h-vowel boundaries,
8. In nasal segments,
9. In obstruent segments,
10. In lateral segments,
11. In h segments.

Pivots can be thus exhaustively classified in 11 groups based on their location within the utterance. Table 7.1 gives the number of tokens in each of the 11 groups for the VCV database.

Group based on location within the utterance	**Number of pivots**
Pivots at vowel-nasal/nasal-vowel (VN/NV) boundaries	103
Pivots at vowel-lateral/lateral-vowel (VL/LV) boundaries	34
Pivots at vowel-obstruent/obstruent-vowel (VO/OV) boundaries	229
Pivots at vowel-semivowel/semivowel-vowel (VS/SV) boundaries	104
Pivots inserted within vowel segments	926
Pivots inserted within semivowel segments	20
Pivots at vowel-h/h-vowel boundaries	25
Pivots in nasal segments	11
Pivots in obstruent segments	22
Pivots in lateral segments	8
Pivots in h segments	1
Total	**1483**

Table 7.1 – Hand-classification of estimated pivots in the VCV database based on their location within the utterance.

Based on the approach suggested in Section 7.1, each pivot will be examined as a sonorant closure, release, or both, depending on its HPROR value. With the proposed approach, pivots that are passed through the nasal detection as possible nasal closures ***and*** releases will effectively increase the number of pivots in the VCV database: each pivot with |HPROR|≤1 at time *t* is practically replaced with two pivots, one with HPROR >1 and one with HPROR <-1 at time *t*. Table 7.2 illustrates the division of pivots based on their HPROR value. The alternative approach of excluding all estimated pivots with |HPROR| ≤1 from the nasal detection analysis would eliminate 2 pivots at nasal boundaries and 110 pivots at non-nasal boundaries or within non-nasal segments. The approach selected in future applications will depend on the user requirements and context.

We are now ready to apply the acoustic criteria for the nasal murmur and nasal boundary on the estimated pivots, and evaluate their performance in terms of:

1. How accurately they detect pivots at nasal boundaries and in nasal segments,
2. How well they reject non-nasal pivots.

We also separate our analysis to the performance of the nasal boundary and nasal murmur criteria.

Group based on location within the utterance	Number of estimated pivots	Pivots with HPROR<-1	Pivots with HPROR>1	Pivots with |HPROR|≤1	Effective number of pivots for nasal detection
Pivots at VN/NV boundaries	103	51	50	2	105
Pivots at VL/LV boundaries	34	16	17	1	35
Pivots at VO/OV boundaries	229	152	72	5	234
Pivots at VS/SV boundaries	104	51	51	2	106
Pivots within vowel segments	926	585	251	90	1016
Pivots within semivowel segments	20	10	8	2	22
Pivots at vowel-h/h-vowel boundaries	25	11	13	1	26
Pivots in nasal segments	11	2	8	1	12
Pivots in obstruent segments	22	4	12	6	28
Pivots in lateral segments	8	3	3	2	10
Pivots in h segments	1	1	0	0	1
Total	**1483**	**886**	**485**	**112**	**1595**

Table 7.2 – Analysis of the pivots based on their application within the nasal detection module. Based on the HPROR value pivots are examined either as potential sonorant closures, releases, or both. Pivots that are examined once as a potential sonorant closure and once as a potential sonorant release effectively increase the number of pivots in the VCV database.

7.4 Performance of the acoustic criteria for the nasal boundary

In this section we present the results of applying the two formulated acoustic criteria for the nasal boundary on the estimated pivots. In the first two sections we evaluate the performance of the ΔED and |ΔH1| acoustic criteria when applied separately and discuss their performance characteristics. In the last section we evaluate their combined effectiveness when detecting pivots at nasal boundaries. In each performance analysis we quote the number of pivots detected as adhering to the requirements of the specific acoustic criterion.

7.4.1 ΔED acoustic criterion

As a reminder to the reader, the ΔED acoustic criterion requires that the energy of the signal across a nasal closure show an increase across a nasal closure and a decline across a nasal release. Because this criterion has a different expectation for a possible closure and release, pivots are passed to this criterion together with their HPROR value that determines how it is applied. Table 7.3 shows the results of applying this criterion to the pivots described in Table 7.2.

Group based on location within the utterance	Number of pivots	Number of pivots detected as nasal	Detection rate (%)
Pivots at VN/NV boundaries	103	103	100.0
Pivots at VL/LV boundaries	35	31	88.6
Pivots at VO/OV boundaries	229	207	90.4
Pivots at VS/SV boundaries	104	98	94.2
Pivots within vowel segments	926	716	77.3
Pivots within semivowel segments	20	12	60.0
Pivots at vowel-h/h-vowel boundaries	25	24	96.0
Pivots in nasal segments	11	8	72.7
Pivots in obstruent segments	22	13	59.1
Pivots in lateral segments	8	5	62.5
Pivots in h segments	1	0	0.0
Total non-nasal pivots	**1380**	**1114**	**80.7**
Total nasal pivots	**103**	**103**	**100**

Table 7.3 – Performance results of applying the ΔED criterion on the 1483 estimated pivots using the approach where pivots with |HPROR| ≤1 are examined twice by the criterion, once as a possible nasal closure and once as a possible nasal release.

The following can be said about the performance characteristics of this acoustic criterion:

- The currently formulated acoustic criterion is designed to minimize the number of missed nasal boundaries. As a consequence, the criterion detects all pivots at true nasal boundaries while allowing a relatively large number of false positives.
- Analysis of the 11 groups of pivots indicates that the criterion has a significantly better rejection rate for pivots inserted within non-nasal segments than at non-nasal boundaries – a possible extension of this criterion could be designed to focus on detecting missing pivots in the signal.
- Because each pivot shows either a net positive or net negative change in the energy of the two bands, ΔED, pivots tested as a potential nasal closure ***and*** release will always have exactly one context pass this acoustic criterion. For example, a pivot at time t, with HPROR = 0 will always be detected as a possible nasal closure or release, but not both.
- While detecting all pivots at true nasal boundaries, this acoustic criterion rejects only about 20% of non-nasal pivots.

Next we evaluate the performance of the |ΔH1| acoustic criterion in the nasal boundary detection.

7.4.2 |ΔH1| acoustic criterion

The |ΔH1| acoustic criterion projects the requirement that energy of the first harmonic be relatively constant across the possible nasal boundary, whether it is a

possible nasal closure or release. Because each pivot is examined in the same manner, independently of its HPROR value, there is no need to pass the pivots' HPROR value to this criterion. The expectation is that this criterion will be only effective at separating true sonorant from the remaining pivots.

Group based on location within the utterance	Number of pivots	Number of pivots detected as nasal	Detection rate (%)
Pivots at VN/NV boundaries	103	102	99.0
Pivots at VL/LV boundaries	35	33	94.3
Pivots at VO/OV boundaries	229	155	67.7
Pivots at VS/SV boundaries	104	99	95.2
Pivots within vowel segments	926	766	82.7
Pivots within semivowel segments	20	18	90.0
Pivots at vowel-h/h-vowel boundaries	25	16	64.0
Pivots in nasal segments	11	11	100
Pivots in obstruent segments	22	21	95.5
Pivots in lateral segments	8	8	100
Pivots in h segments	1	1	100
Total non-nasal pivots	**1380**	**1128**	**81.7**
Total nasal pivots	**103**	**102**	**99.0**

Table 7.4 – Performance results of applying the |ΔH1| acoustic criterion on the 1483 estimated pivots. The criterion is applied uniformly across all estimated pivots, independently of their HPROR value.

The following is a summary of the performance characteristics for the |ΔH1| acoustic criterion:

- The criterion shows a high detection rate for pivots at all sonorant boundaries, such as laterals, semivowels, and nasals – this aligns with the theoretical

expectation that small change in the pressure at the glottis for sonorant segments will result in a relatively constant signal at low frequencies.

- The single rejected pivot at a true nasal boundary is located at the time of the nasal release in the utterance [ɛŋɛ].The COLEA f0track appears consistent and accurate across pivots and utterances.
- Pivots located within vowel segments that are rejected with this criterion are located either at the beginning or end of voicing in an utterance.
- This acoustic criterion is formulated with the goal of minimizing the number of missed nasal boundaries, while allowing for a high rate of false positives – consequently only about 20% of non-nasal pivots are rejected by this criterion.

Lastly we combine the formulated acoustic criteria in a test for the nasal boundary.

7.4.3 Combined performance of the nasal boundary criteria

The difference in the nature of the formulated acoustic criteria suggests that their combined use will show significant improvement with regard to either of the isolated performance results. On the one hand, the ΔED acoustic criterion has been found to discriminate well between pivots at boundaries and within segments: the majority of pivots that met this requirement indicated some abrupt acoustic change, being either located at vowel-consonant boundaries or at the beginning or end of voicing within vowel segments. The |ΔH1| acoustic criterion, on the other hand, imposed a requirement that pivot sees almost no change in the energy at low frequencies, thus effectively

eliminating only non-sonorant boundaries. In this section we show the results of applying the combined criteria to the estimated pivots.

Group based on location within the utterance	Number of pivots	Number of pivots detected as nasal	Detection rate (%)
Pivots at VN/NV boundaries	103	102	99.0
Pivots at VL/LV boundaries	35	31	88.6
Pivots at VO/OV boundaries	229	137	59.8
Pivots at VS/SV boundaries	104	88	84.6
Pivots within vowel segments	926	598	64.6
Pivots within semivowel segments	20	10	50.0
Pivots at vowel-h/h-vowel boundaries	25	15	60.0
Pivots in nasal segments	11	8	72.7
Pivots in obstruent segments	22	12	54.5
Pivots in lateral segments	8	5	62.5
Pivots in h segments	1	0	0.0
Total non-nasal pivots	**1380**	**904**	**65.5**
Total nasal pivots	**103**	**102**	**99.0**

Table 7.5 – Performance results of applying the combined acoustic criteria for the nasal boundary on the 1483 estimated pivots. The $|\Delta H1|$ acoustic criterion is applied uniformly to all pivots independently of their HPROR value, while the ΔED criterion required the HPROR value to apply the correct expectation to each pivot.

As in the previous two sections, the following observations can be made regarding the performance of the combined acoustic criteria:

- The performance is characterized by the high detection rate of both true nasal and non-nasal pivots. More than 60% of all estimated non-nasal pivots are detected as

showing properties of the nasal boundary. Detection rates for pivots at lateral and semivowel boundaries are highest among the non-nasal pivots.

- The criteria appear to be especially ill-suited for databases with a large number of semivowel and lateral boundaries – these contexts will see a significant degradation in the performance of the criteria due to the high rate of false positives.
- The large number of false positives shows that a larger set of more effective acoustic criteria is needed in order to efficiently separate nasal from non-nasal boundaries if the goal of the algorithm is still to minimize the miss rate with the same decision process.
- One pivot in the vowel-nasal group that is not detected by the combined criteria is located at the time of the nasal release in the utterance [ɛŋɛ] and discussed in Section 7.4.2.
- Because they analyze the region in the immediate vicinity of the estimated pivot, these criteria will not depend on the duration of the vowel and will show limited dependence on the duration of the nasal murmur (all calculations are limited to the time interval of about 30ms on either side of the estimated pivot).

We next turn our attention to the formulated acoustic criteria for the nasal murmur.

7.5 Performance of the acoustic criteria for the nasal murmur

In Chapter 5 we reconstructed Chen's algorithm for the nasal murmur detection and adjusted its criteria by suggesting a new structure and limits for the specific acoustic

cues. As a reminder to the reader, the algorithm used six acoustic cues to examine the properties of signal energy and spectral tilt, and compare them against the quantified expectation for the nasal segments. From the computational point of view, the six cues are calculated almost simultaneously, making it computationally rather expensive to apply each criterion in isolation. The majority of the six cues, in addition, impose a requirement on the spectral characteristics of some frequency band with respect to the lowest 0-788 Hz band – this further complicates the applicability of formulated acoustic criteria in isolation from each other. For these reasons Table 7.6 shows the performance characteristics of the aggregate nasal murmur criteria when applied to the 1483 estimated pivots in this VCV database. Pivots are examined with the approach illustrated in Figure 7.2 that relies on the HPROR value to determine whether the pivot is a likely nasal closure or release.

The results in Table 7.6 indicate that:

- The nasal murmur criteria discriminate poorly between nasals and the remaining sonorant consonants – about 63% of lateral and semivowel pivots are detected as showing characteristics of the nasal murmur.
- Approximately 20% of pivots within vowel segments that were detected as showing characteristics of the nasal murmur were adjacent to the nasal segment. The proximity of the pivot to the nasal boundary causes most measurements in the nasal murmur algorithm to be taken in the region of the signal that actually belongs to the nasal segment. As such, these pivots are actually measured correctly and their detection in the nasal murmur algorithm is advantageous in possible further acoustic analysis.

Group based on location within the utterance	Number of pivots	Number of pivots detected as nasal	Detection rate (%)
Pivots at VN/NV boundaries	103	102*	99.0
Pivots at VL/LV boundaries	35	22	62.9
Pivots at VO/OV boundaries	229	42	18.3
Pivots at VS/SV boundaries	104	69	66.3
Pivots within vowel segments	926	155	16.7
Pivots within semivowel segments	20	9	45.0
Pivots at vowel-h/h-vowel boundaries	25	0	0.0
Pivots in nasal segments	11	9	81.8
Pivots in obstruent segments	22	9	40.9
Pivots in lateral segments	8	5	62.5
Pivots in h segments	1	0	0
Total non-nasal pivots	**1380**	**320**	**23.3**
Total nasal pivots	**103**	**102**	**99.0**

Table 7.6 – Performance results of the aggregate criteria for the nasal murmur applied to the 1483 estimated pivots. The criteria use the HPROR value associated with each pivot to determine whether it should be examined as a possible nasal closure, release, or both.

*** One pivot at a true nasal boundary with $|HPROR| \leq 1$ was detected as two nasal pivots when examined as a possible sonorant closure and release. We considered this pivot as contributing a single token to the Number of pivots detected as nasal column.**

- The single pivot at a true nasal boundary that was detected as not showing properties of the nasal murmur was located at the nasal closure in utterance [ɑŋɑ]. The pivot was rejected on the basis of not meeting the A_1-A_2 requirement in the first measured frame, closest to the boundary. The remaining frames adhered to all formulated criteria. This result

indicates that further testing of the criteria will either modify the proposed thresholds or the structure of the decision process.

- Analysis of the two pivots inserted within nasal segment that were detected as not showing characteristics of the nasal murmur indicates that:
    - One pivot failed because its proximity to the vowel-nasal boundary resulted in the large fluctuations of the RMS value, which in term caused all but one frame to be discarded in the nasal murmur algorithm. The requirement that at least two frames comply with the required acoustic criteria caused this pivot to be rejected for the nasal murmur characteristics.
    - Rejection of the second pivot was also caused by its proximity to the vowel boundary. This time, however, the initial calculations across the boundary caused the initial frames not to pass the RMS criterion and thus not be examined for the remaining acoustic characteristics. Once in the vowel segment, however, the RMS value became stable while the largest peak in the lowest frequency band aligned with the true first formant. Failure of this cue was specifically caused by the f1 value in these frames.

The significance of these two cases is in recognizing contexts that may limit the algorithm's functionality in spontaneous speech, where nasal segments can measure less than 50ms in duration from the estimated landmark point. With such nasal segments, most measurements will be

made in the vicinity of the nasal boundary and may require reformulation of the decision process.

7.6 Combined performance of the nasal boundary and murmur criteria

In American English at least one segment adjacent to the nasal is a vowel. It is therefore significant to evaluate the performance of the nasal boundary and nasal murmur criteria when applied in combination on the estimated pivots. Table 7.7 gives the performance results.

Group based on location within the utterance	Number of pivots	Number of pivots detected as nasal	Detection rate (%)
Pivots at VN/NV boundaries	103	101	98.1
Pivots at VL/LV boundaries	35	19	54.3
Pivots at VO/OV boundaries	229	23	10.0
Pivots at VS/SV boundaries	104	58	55.8
Pivots within vowel segments	926	115	12.4
Pivots within semivowel segments	20	4	25.0
Pivots at vowel-h/h-vowel boundaries	25	0	0.0
Pivots in nasal segments	11	7	63.6
Pivots in obstruent segments	22	0	0.0
Pivots in lateral segments	8	1	12.5
Pivots in h segments	1	0	0.0
Total non-nasal pivots	**1380**	**227**	**16.4**
Total nasal pivots	**103**	**101**	**98.1**

Table 7.7 – Results of applying the combined criteria for the nasal boundary and nasal murmur on the 1483 estimated pivots in the VCV database.

If each nasal pivot is also expected to be a vowel-nasal boundary, as in a VCV database, application of the combined acoustic criteria for the nasal boundary and nasal murmur will show the following characteristics:

- A large majority of pivots at true nasal boundaries will be detected by the combined acoustic criteria – for the VCV database this detection rate is 98.1%.
- A majority of the false positives will belong to pivots at lateral and semivowel boundaries, and within nasal segments. Additional acoustic cues need to be examined in the future that will specifically target laterals and semivowels.
- Most pivots within vowel segments that adhere to both criteria are located close to the nasal, lateral, or semivowel boundary. With further processing this information can be used to possibly compensate for the missing pivots.
- For databases that may be more heavily weighted towards laterals and semivowels we will see a significant degradation in the performance of the criteria.
- Depending on the requirements, it is possible to formulate the criteria using the maximum likelihood ratio test or Neyman-Pearson detection principles that will further optimize its performance. Our reasoning behind choosing to maximize the detection rate without a specified limit on the rate of false positives is that information from the lexicon and other modules within the LAFF system will help eliminate some of the false positives. No mechanism currently exists that would compensate for the missing nasal pivots, making them rather costly for the system.

7.7 Minimizing the computational power by the algorithm

In Chapter 2 we recommended that acoustic criteria for nasality be ordered in such a way as to minimize the required computational power when examining estimated pivots for nasal characteristics. During the algorithm design, however, the most optimal solution suggested that all nasal murmur criteria be calculated almost simultaneously, thus not allowing any ordering or separation. Our only suggestion in minimizing the overall computational requirement is in using the nasal boundary and nasal murmur criteria in combination. If used in combination, using the ΔED acoustic criterion first will determine whether pivots with $|\text{HPROR}| \leq 1$ should be passed as possible nasal closures or releases to the next algorithm. Taking this approach would eliminate 112 of 1483 estimated pivots from further consideration.

7.8 Contributions and future work

This thesis study is the first step in the creation of an automated nasal detection module in a feature-based system. At the beginning of the project we examined the consonant landmark estimation results and their influence on the design of the nasal module. We then used the estimated landmarks to design and test algorithms for various acoustic cues until their results conformed to the theoretical expectations. In the cases where theoretical expectations were not met, we attempted to identify possible reasons and make a recommendation regarding how they can be solved in the future work. Consistencies among a number of acoustic cues for nasal sonorant landmarks were translated into acoustic criteria that were used to separate nasals from non-nasal landmarks. The criteria defined in this study are not optimized to yield most efficient

classification; the automated algorithms developed to extract eight promising cues, however, allow for a relatively quick and consistent analysis of further contexts and speakers, and further analysis of decision structures that would yield more robust and accurate performance results. With a simple formulation of the criteria based on the preliminary analysis of estimated sonorant landmarks and observations made in past studies, we detected 98.2% of true nasal pivots from the CLD output.

Based on the work completed in this thesis, there are three possible directions for future work. The three directions are:

1. Expanding the existing set of acoustic criteria to better address the task of rejecting non-nasal pivots at semivowel and lateral segments,
    a. In Chapter 4 we suggested measuring the energy difference between the 0-350Hz and 350-2000Hz band. Including higher frequencies than tested in this study guarantees to capture the introduction of the nasal antiformant in the 800-2000Hz region.
    b. Preliminary analysis of the ED cue measured in the portion of the signal following a nasal closure or preceding a nasal release showed promising results when tested with the ANOVA statistical analysis tool, though this cue was not included in the current formulation of the criteria.
2. Collecting more observations by running the same algorithms on a variety of databases, and analyzing the formulation of acoustic criteria that yields highest accuracy and robustness using statistical classifiers,

a. The advantage of using statistical classifiers rather than firm thresholds is that classification of each pivot is accompanied by a confidence rating that is advantageous in further acoustic processing. For example, if a set of acoustic cues classify a pivot as nasal with a 58% and lateral with 34% confidence rating, this information would be advantageous when accessing the lexicon to stipulate a possible cohort of words. For each member of cohort, a module could be made to do an "internal synthesis" of the expected parameters and compare with the observations made in the nasal module. With an access to the lexicon, each cohort would also identify the vowel adjacent to the nasal – this information in turn, would allow us to better estimate whether the pivot is nasal or not.

3. Porting the algorithms to spontaneous speech and observing the change in the number of available cues and their behavior in obscured environments. Measurements made for these databases will most likely reformulate the way in which measured cues are used during the decision process.

The main contribution of this study is in formulating a set of eight acoustic cues in terms of automated algorithms that can be used within the nasal module or included in other analysis. Measurements made these algorithms will guarantee a consistent method of examining large volumes of databases and contexts. Information collected in such a

way will allow us to define more general and robust acoustic criteria for the nasal detection, and shed more light on possible extensions to the nasal detection module.

References

[1] Chen, M.Y. (submitted) *Nasal Detection module for a Knowledge-based Speech Recognition System*

[2] Chen, M.Y. (1995) *Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers*, Journal of the Acoustical Society of America, 98 (5), 2443-2453

[3] Chen, M.Y. (1997) *Acoustic correlates of English and French nasalized vowels*, Journal of the Acoustical Society of America, 102 (4), 2360-2370

[4] Chen, M.Y. (2000) *Nasal Detection module for a knowledge-based Speech Recognition System*, appears in Proc. 6th International Conference on Spoken Language Processing (ICSLP2000) in Beijing, China on October 16-20, 2000.

[5] Chen, M.Y. (submitted) *Modifications of the Nasal Detection Module for a Knowledge-based Speech Recognition System*

[6] Das, S., Bakis, R., Nadas, A., Nahamoo, D. & Picheny, M. *Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system*, Proc. of the IEEE International Conference on Acoustical Speech Signal Processing, 1993, 2, 71-74.

[7] Devore, J.L. *Probability and Statistics for Engineering and the Sciences*, 4th ed. Wadsworth Publishing, 1995.

[8] Glass, J.R. (1984) *Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment*, MS and EE thesis, Massachusetts Institute of Technology, Cambridge, MA

[9] Glass, J. R. and Zue, V. W. (1984) *An Acoustic Study of Nasal Consonants in American English*, presented at the 108th Meeting of the Acoustical Society of America in Minneapolis, Minnesota

[10] Halle, M. (1992) Features. In *Oxford international encyclopedia of linguistics*, New York: Oxford University Press.

[11] Hattori, S. and Fujimura, O. (1958) *Nasalization of Vowels in Relation to Nasals*, Journal of the Acoustical Society of America, 30, 267-274

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

[13] Liu, S.A. (1995) *Landmark Detection for Distinctive Feature-Based Speech Recognition*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA

[14] Liu, S.A. (1996) *Landmark detection for distinctive feature-based speech recognition*, Journal of the Acoustical Society of America, 100, 3417-3430

[15] Maeda, S. (1982) *Acoustic cues of vowel nasalization: a simulation study*, Journal of the Acoustical Society of America, 72, S102

[16] Mermelstein, P. (1977) *On detecting nasals in continuous speech*, Journal of the Acoustical Society of America, 61, 581-587

[17] Nakata, K. (1959) *Synthesis and Perception of Nasal Consonants*, Journal of the Acoustical Society of America, 31 (6), 661-666

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

[19] Stevens, K.N. (1969) *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. 1971

[20] Stevens, K.N., Kalikow, D.N., Willemain, T.R. (1975) *A Miniature Accelerometer for Detecting Glottal Waveforms and Nasalization*, Journal of Speech and Hearing Research, Vol. 18, No. 3

[21] Stevens, K.N., Andrade, A., Viana, M.C. (1988) *Perception of vowel nasalization in VC contexts: A cross-language study*, ASA Miami

[22] Stevens, K.N. (1990) *Lexical Access from Features*, Written version of a talk given at the workshop on Speech Technology for Man-Machine Interaction, held at the Tata Institute of Fundamental Research in Bombay, India

[23] Stevens, K.N. (1998) *Acoustic Phonetics*. Cambridge, MA: MIT Press.

[24] Stevens, K.N. (2002) *Toward a model for lexical access based on acoustic landmarks and distinctive features*, Journal of the Acoustical Society of America, 111, 1872-1891

Appendix A: Results of the Sonorant Landmark Estimation

Results of the sonorant landmark estimation on the VCV database of 453 utterances with Liu's CLD. The 251 estimated sonorant landmarks are hand-classified into one of six groups:

1. Landmarks at vowel-nasal/nasal-vowel boundary,
2. Landmarks at vowel-lateral/lateral-vowel boundary,
3. Landmarks within vowel segments,
4. Landmarks at vowel-semivowel/semivowel-vowel boundary,
5. Landmarks at vowel-obstruent/obstruent-vowel boundary,
6. Landmarks in semivowel segments.

Utterance	Landmark	Time	Comment	Hand-classification
aaaa_cb	----------	----------	----------	
aaaa_dw	----------	----------	----------	
aaaa_ks	----------	----------	----------	
aabaa_cb	----------	----------	----------	
aabaa_dw	----------	----------	----------	
aabaa_ks	----------	----------	----------	
aachaa_cb	----------	----------	----------	
aachaa_dw	----------	----------	----------	
aachaa_ks	----------	----------	----------	
aadaa_cb	-s	188	closure for d	vowel-obstruent
aadaa_dw	-s	231	closure for d	vowel-obstruent
aadaa_ks	-s	242	closure for d	vowel-obstruent
aadhaa_cb	----------	----------	----------	
aadhaa_dw	+s	530	release for dh	obstruent-vowel
aadhaa_ks	----------	----------	----------	
aadjaa_cb	----------	----------	----------	
aadjaa_dw	-s	345	closure for dj	vowel-obstruent
aadjaa_ks	----------	----------	----------	
aafaa_cb	----------	----------	----------	
aafaa_dw	----------	----------	----------	
aafaa_ks	----------	----------	----------	
aagaa_cb	----------	----------	----------	
aagaa_dw	-s	249	closure for g	vowel-obstruent
aagaa_ks	-s	252	closure for g	vowel-obstruent
aahaa_cb	----------	----------	----------	
aahaa_dw	----------	----------	----------	
aakaa_ks	----------	----------	----------	
aalaa_cb	-s	215	closure for l	vowel-lateral
	+s	375	release for l	lateral-vowel
aalaa_dw	----------	----------	----------	
aalaa_ks	+s	338	release for l	lateral-vowel
aamaa_cb	----------	----------	----------	
aamaa_dw	-s	274	closure for m	vowel-nasal

Utterance	Landmark	Time	Comment	Hand-classification
	+s	427	release for m	nasal-vowel
aamaa_ks	+s	63	beg of V1	within vowel
	-s	206	closure for m	vowel-nasal
	+s	328	release for m	nasal-vowel
aanaa_cb	-s	197	closure for n	vowel-nasal
	+s	331	release for n	nasal-vowel
aanaa_dw	-s	304	closure for n	vowel-nasal
	+s	471	release for n	nasal-vowel
aanaa_ks	-s	210	closure for n	vowel-nasal
	+s	325	release for n	nasal-vowel
aangaa_cb	----------	----------	----------	
aangaa_dw	-s	255	closure for ng	vowel-nasal
	+s	452	release for ng	nasal-vowel
aangaa_ks	-s	217	closure for ng	vowel-nasal
	+s	349	release for ng	nasal-vowel
aapaa_cb	----------	----------	----------	
aapaa_dw	----------	----------	----------	
aapaa_ks	----------	----------	----------	
aaraa_cb	----------	----------	----------	
aaraa_dw	----------	----------	----------	
aaraa_ks	----------	----------	----------	
aasaa_cb	----------	----------	----------	
aasaa_dw	----------	----------	----------	
aasaa_ks	----------	----------	----------	
aashaa_cb	----------	----------	----------	
aashaa_dw	----------	----------	----------	
aashaa_ks	+s	74	beg of vowel	within vowel
aataa_cb	----------	----------	----------	
aataa_dw	----------	----------	----------	
aataa_ks	----------	----------	----------	
aathaa_cb	----------	----------	----------	
aathaa_dw	+s	100	beg of V1	within vowel
aathaa_ks	----------	----------	----------	
aavaa_cb	----------	----------	----------	
aavaa_dw	-s	720	end of V2	within vowel
aavaa_ks	----------	----------	----------	
aawaa_cb	----------	----------	----------	
aawaa_dw	----------	----------	----------	
aawaa_ks	+s	371	release for w	semivowel-vowel
aayaa_cb	+s	379	release for y	semivowel-vowel
aayaa_dw	----------	----------	----------	
aayaa_ks	----------	----------	----------	
aazaa_cb	-s	193	closure for z	vowel-obstruent
aazaa_dw	+s	87	beg of utt.	within vowel
aazaa_ks	----------	----------	----------	
aazhaa_cb	----------	----------	----------	

Utterance	Landmark	Time	Comment	Hand-classification
aazhaa_dw	----------	----------	----------	
aazhaa_ks	----------	----------	----------	
ahah_cb	----------	----------	----------	
ahah_dw	----------	----------	----------	
ahah_ks	----------	----------	----------	
ahbah_cb	----------	----------	----------	
ahbah_dw	----------	----------	----------	
ahbah_ks	+s	76	beg of V1	within vowel
ahchah_cb	----------	----------	----------	
ahchah_dw	----------	----------	----------	
ahchah_ks	----------	----------	----------	
ahdah_cb	----------	----------	----------	
ahdah_dw	-s	277	closure for d	vowel-obstruent
ahdah_ks	-s	196	closure for d	vowel-obstruent
ahdhah_cb	+s	55	beg of utt.	within vowel
	+s	301	release for dh	obstruent-vowel
ahdhah_dw	----------	----------	----------	
ahdhah_ks	----------	----------	----------	
ahdjah_cb	----------	----------	----------	
ahdjah_dw	-s	208	closure for dj	vowel-obstruent
ahdjah_ks	-s	174	closure for dj	vowel-obstruent
ahfah_cb	----------	----------	----------	
ahfah_dw	----------	----------	----------	
ahfah_ks	----------	----------	----------	
ahgah_cb	----------	----------	----------	
ahgah_dw	-s	258	closure for g	vowel-obstruent
ahgah_ks	-s	194	closure for g	vowel-obstruent
ahhah_cb	----------	----------	----------	
ahhah_dw	----------	----------	----------	
ahhah_ks	-s	542	end of utt.	within vowel
ahkah_cb	----------	----------	----------	
ahkah_dw	----------	----------	----------	
ahkah_ks	----------	----------	----------	
ahlah_cb	+s	427	release for l	lateral-vowel
ahlah_dw	-s	134	mid vowel	within vowel
ahlah_ks	+s	344	release for l	lateral-vowel
ahmah_cb	-s	196	closure for m	vowel-nasal
	+s	330	release for m	nasal-vowel
ahmah_dw	+s	447	release for m	nasal-vowel
ahmah_ks	-s	184	closure for m	vowel-nasal
	+s	323	release for m	nasal-vowel
ahnah_cb	-s	169	closure for n	vowel-nasal
	+s	329	release for n	nasal-vowel
ahnah_dw	-s	216	closure for n	vowel-nasal
	+s	487	release for n	nasal-vowel
ahnah_ks	-s	184	closure for n	vowel-nasal

Utterance	Landmark	Time	Comment	Hand-classification
	+s	334	release for n	nasal-vowel
ahngah_cb	-s	169	closure for n	vowel-nasal
	+s	359	release for n	nasal-vowel
ahngah_dw	+s	459	release for n	nasal-vowel
ahngah_ks	+s	57	beg of utt.	within vowel
ahpah_cb	----------	----------	----------	
ahpah_dw	----------	----------	----------	
ahpah_ks	----------	----------	----------	
ahrah_cb	+s	430	release for r	semivowel-vowel
ahrah_dw	----------	----------	----------	
ahrah_ks	----------	----------	----------	
ahsah_cb	----------	----------	----------	
ahsah_dw	----------	----------	----------	
ahsah_ks	----------	----------	----------	
ahshah_cb	----------	----------	----------	
ahshah_dw	----------	----------	----------	
ahshah_ks	----------	----------	----------	
ahtah_cb	----------	----------	----------	
ahtah_dw	----------	----------	----------	
ahtah_ks	----------	----------	----------	
ahthah_cb	----------	----------	----------	
ahthah_dw	----------	----------	----------	
ahthah_ks	----------	----------	----------	
ahvah_cb	-s	233	closure for v	vowel-obstruent
	+s	371	release for v	obstruent-vowel
ahvah_dw	----------	----------	----------	
ahvah_ks	----------	----------	----------	
ahwah_cb	----------	----------	----------	
ahwah_dw	----------	----------	----------	
ahwah_ks	----------	----------	----------	
ahyah_cb	----------	----------	----------	
ahyah_dw	----------	----------	----------	
ahyah_ks	-s	192	closure for y	vowel-semivowel
ahzah_cb	----------	----------	----------	
ahzah_dw	----------	----------	----------	
ahzah_ks	----------	----------	----------	
ahzhah_cb	----------	----------	----------	
ahzhah_dw	----------	----------	----------	
ahzhah_ks	-s	553	end of utt.	within vowel
ehbeh_cb	----------	------	----------	
ehbeh_dw	-s	197	closure for b	vowel-obstruent
	-s	482	end of V2	within vowel
ehbeh_ks	-s	235	closure for b	vowel-obstruent

Utterance	Landmark	Time	Comment	Hand-classification
ehcheh_cb	-s	565	end of utt.	within vowel
ehcheh_dw	----------	----------	----------	
ehcheh_ks	----------	------	----------	
ehdeh_cb	-s	140	closure for d	vowel-obstruent
ehdeh_dw	-s	576	end of V2	within vowel
ehdeh_ks	-s	223	closure for d	vowel-obstruent
ehdheh_cb	----------	------	----------	
ehdheh_dw	----------	----------	----------	
ehdheh_ks	+s	126	mid V1	within vowel
ehdjeh_cb	-s	147	closure for dj	vowel-obstruent
ehdjeh_dw	-s	97	mid V1	within vowel
	-s	205	closure for dj	vowel-obstruent
ehdjeh_ks	-s	216	closure for dj	vowel-obstruent
eheh_cb	----------	------	----------	
eheh_dw	----------	------	----------	
eheh_ks	----------	----------	----------	
ehfeh_cb	----------	------	----------	
ehfeh_dw	----------	------	----------	
ehfeh_ks	----------	----------	----------	
ehgeh_cb	-s	148	closure for g	vowel-obstruent
	-s	511	end of V2	within vowel
ehgeh_dw	----------	------	----------	
ehgeh_ks	-s	194	closure for g	vowel-obstruent
ehheh_cb	----------	------	----------	
ehheh_dw	----------	----------	----------	
ehheh_ks	-s	565	end of utt.	within vowel
ehkeh_cb	----------	----------	----------	
ehkeh_dw	----------	------	----------	
ehkeh_ks	----------	------	----------	
ehleh_cb	+s	368	release for l	lateral-vowel
ehleh_dw	----------	----------	----------	
ehleh_ks	+s	89	beg of utt.	within vowel
	+s	325	release for l	lateral-vowel
ehmeh_cb	-s	192	closure for m	vowel-nasal
	+s	359	release for m	nasal-vowel
ehmeh_dw	-s	209	closure for m	vowel-nasal
	+s	394	release for m	nasal-vowel
ehmeh_ks	-s	174	closure for m	vowel-nasal
	+s	306	release for m	nasal-vowel
	-s	485	end of utt.	within vowel
ehneh_cb	-s	147	closure for n	vowel-nasal
	+s	345	release for n	nasal-vowel
	-s	611	end of utt.	within vowel
ehneh_dw	-s	206	closure for n	vowel-nasal
	+s	413	release for n	nasal-vowel
	-s	582	mid V2	within vowel

Utterance	Landmark	Time	Comment	Hand-classification
	+s	628	end of V2	within vowel
ehneh_ks	-s	206	closure for n	vowel-nasal
	+s	334	release for n	nasal-vowel
ehngeh_cb	-s	200	closure for ng	vowel-nasal
ehngeh_dw	-s	235	closure for ng	vowel-nasal
	+s	424	release for ng	nasal-vowel
ehngeh_ks	-s	197	closure for ng	vowel-nasal
ehpeh_cb	-s	497	end of utt.	within vowel
ehpeh_dw	+s	365	release for p	obstruent-vowel
ehpeh_ks	-s	525	end of utt.	within vowel
ehreh_cb	----------	----------	----------	
ehreh_dw	----------	----------	----------	
ehreh_ks	-s	553	end of utt.	within vowel
ehseh_cb	----------	------	----------	
ehseh_dw	----------	----------	----------	
ehseh_ks	+s	79	beg of V1	within vowel
ehsheh_cb	----------	------	----------	
ehsheh_dw	----------	----------	----------	
ehsheh_ks	-s	191	closure for sh	vowel-obstruent
ehteh_cb	----------	------	----------	
ehteh_dw	----------	------	----------	
ehteh_ks	----------	----------	----------	
ehtheh_cb	-s	547	end of utt.	within vowel
ehtheh_dw	----------	------	----------	
ehtheh_ks	----------	------	----------	
ehveh_cb	----------	----------	----------	
ehveh_dw	----------	------	----------	
ehveh_ks	+s	297	release for v	obstruent-vowel
ehweh_cb	-s	238	closure for w	vowel-semivowel
ehweh_dw	+s	392	release for w	semivowel-vowel
ehweh_ks	----------	----------	----------	
ehyeh_cb	----------	----------	----------	
ehyeh_dw	----------	----------	----------	
ehyeh_ks	----------	----------	----------	
ehzeh_cb	----------	------	----------	
ehzeh_dw	----------	----------	----------	
ehzeh_ks	-s	206	closure for z	vowel-obstruent
	-s	534	end of utt.	within vowel
ehzheh_cb	----------	------	----------	
ehzheh_dw	-s	337	closure for zh	vowel-obstruent
	-s	653	mid V2	within vowel
ehzheh_ks	----------	----------	----------	
iybiy_cb	-s	553	end of utt.	within vowel

Utterance	Landmark	Time	Comment	Hand-classification
iybiy_dw	-s	533	end of V2	within vowel
iybiy_ks	+s	67	beg of utt.	within vowel
iychiy_cb	----------	----------	----------	
iychiy_dw	-s	99	beg of V1	within vowel
	-s	624	mid V2	within vowel
iychiy_ks	----------	----------	----------	
iydhiy_cb	----------	----------	----------	
iydhiy_dw	----------	----------	----------	
iydhiy_ks	----------	----------	----------	
iydiy_cb	----------	----------	----------	
iydiy_dw	-s	215	closure for d	vowel-obstruent
iydiy_ks	-s	197	closure for d	vowel-obstruent
iydjiy_cb	-s	165	closure for dj	vowel-obstruent
iydjiy_dw	-s	333	closure for dj	vowel-obstruent
iydjiy_ks	-s	599	end of utt.	within vowel
iyfiy_cb	-s	581	end of utt.	within vowel
iyfiy_dw	+s	102	beg of utt.	within vowel
iyfiy_ks	-s	579	end of utt.	within vowel
iygiy_cb	----------	----------	----------	
iygiy_dw	-s	252	closure for g	vowel-obstruent
iygiy_ks	+s	81	beg of V1	within vowel
iyhiy_cb	----------	----------	----------	
iyhiy_dw	----------	----------	----------	
iyhiy_ks	----------	----------	----------	
iyiy_cb	----------	----------	----------	
iyiy_dw	----------	----------	----------	
iyiy_ks	-s	580	end of utt.	within vowel
iykiy_cb	-s	500	end of utt.	within vowel
iykiy_dw	-s	527	end of V2	within vowel
iykiy_ks	+s	53	beg of utt.	within vowel
iyliy_cb	-s	190	closure for l	vowel-lateral
	+s	327	release for l	lateral-vowel
iyliy_dw	----------	----------	----------	
iyliy_ks	-s	205	closure for l	vowel-lateral
iymiy_cb	-s	187	closure for m	vowel-nasal
	+s	294	release for m	nasal-vowel
iymiy_dw	-s	566	mid V2	within vowel
iymiy_ks	-s	216	closure for m	vowel-nasal
	+s	339	release for m	nasal-vowel
	-s	555	end of utt.	within vowel
iyngiy_cb	-s	129	closure for ng	vowel-nasal
iyngiy_dw	-s	177	closure for ng	vowel-nasal
iyngiy_ks	**missing**	**in the**	**database**	
iyniy_cb	-s	190	closure for n	vowel-nasal
	+s	326	release for n	nasal-vowel
iyniy_dw	-s	269	closure for n	vowel-nasal

Utterance	Landmark	Time	Comment	Hand-classification
	+s	472	release for n	nasal-vowel
iyniy_ks	-s	201	closure for n	vowel-nasal
	+s	334	release for n	nasal-vowel
	-s	569	end of utt.	within vowel
iypiy_cb	-s	503	end of utt.	within vowel
iypiy_dw	-s	552	end of V2	within vowel
iypiy_ks	-s	540	end of utt.	within vowel
iyriy_cb	----------	----------	----------	
iyriy_dw	-s	595	mid V2	within vowel
iyriy_ks	-s	578	end of utt.	within vowel
iyshiy_cb	----------	----------	----------	
iyshiy_dw	----------	----------	----------	
iyshiy_ks	-s	148	closure for sh	vowel-obstruent
iysiy_cb	-s	582	end of utt.	within vowel
iysiy_dw	----------	----------	----------	
iysiy_ks	----------	----------	----------	
iythiy_cb	-s	591	end of utt.	within vowel
iythiy_dw	----------	----------	----------	
iythiy_ks	-s	615	end of utt.	within vowel
iytiy_cb	-s	510	end of utt.	within vowel
iytiy_dw	----------	----------	----------	
iytiy_ks	----------	----------	----------	
iyviy_cb	-s	377	mid V2	within vowel
	-s	515	end of V2	within vowel
iyviy_dw	----------	----------	----------	
iyviy_ks	-s	516	end of utt.	within vowel
iywiy_cb	-s	776	end of utt.	within vowel
iywiy_dw	+s	404	release for w	semivowel-vowel
	-s	550	end of V2	within vowel
iywiy_ks	-s	189	closure for w	vowel-semivowel
iyyiy_cb	----------	----------	----------	
iyyiy_dw	-s	649	end of V2	within vowel
iyyiy_ks	+s	351	release for y	semivowel-vowel
iyzhiy_cb	----------	----------	----------	
iyzhiy_dw	-s	101	beg of V1	within vowel
	-s	206	mid V1	within vowel
iyzhiy_ks	----------	----------	----------	
iyziy_cb	----------	----------	----------	
iyziy_dw	----------	----------	----------	
iyziy_ks	----------	----------	----------	
owbow_cb	----------	----------	----------	
owbow_dw	----------	----------	----------	
owbow_ks	----------	----------	----------	

Utterance	Landmark	Time	Comment	Hand-classification
owchow_cb	----------	----------	----------	
owchow_dw	----------	----------	----------	
owchow_ks	----------	----------	----------	
owdhow_cb	-s	230	closure for dh	vowel-obstruent
	+s	337	release for dh	obstruent-vowel
owdhow_dw	-s	301	closure for dh	vowel-obstruent
owdhow_ks	----------	----------	----------	
owdjow_cb	----------	----------	----------	
owdjow_dw	----------	----------	----------	
owdjow_ks	+s	322	release for dj	obstruent-vowel
owdow_cb	----------	----------	----------	
owdow_dw	+s	107	beg of utt.	within vowel
owdow_ks	-s	214	closure for d	within vowel
owfow_cb	----------	----------	----------	
owfow_dw	-s	171	closure for f	vowel-obstruent
	-s	590	end of V2	within vowel
owfow_ks	----------	----------	----------	
owgow_cb	-s	605	end of utt.	within vowel
owgow_dw	----------	----------	----------	
owgow_ks	----------	----------	----------	
owhow_cb	----------	----------	----------	
owhow_dw	-s	626	end of V2	within vowel
owhow_ks	----------	----------	----------	
owkow_cb	-s	542	end of utt.	within vowel
owkow_dw	----------	----------	----------	
owkow_ks	----------	----------	----------	
owlow_cb	+s	364	release for l	lateral-vowel
	-s	692	end of utt.	within vowel
owlow_dw	+s	439	release for l	lateral-vowel
owlow_ks	+s	354	release for l	lateral-vowel
owmow_cb	+s	354	release for m	nasal-vowel
	-s	677	end of utt.	within vowel
owmow_dw	+s	475	release for m	nasal-vowel
owmow_ks	+s	301	release of m	nasal-vowel
owngow_cb	----------	----------	----------	
owngow_dw	+s	439	release for ng	nasal-vowel
owngow_ks	+s	320	release for ng	nasal-vowel
ownow_cb	+s	367	release for n	nasal-vowel
ownow_dw	+s	91	beg of utt.	within vowel
	-s	311	closure for n	vowel-nasal
ownow_ks	-s	214	closure for n	vowel-nasal
	+s	328	release for n	nasal-vowel
owow_cb	----------	----------	----------	
owow_dw	----------	----------	----------	
owow_ks	----------	----------	----------	
owpow_cb	-s	553	end of utt.	within vowel

Utterance	Landmark	Time	Comment	Hand-classification
owpow_dw	----------	----------	----------	
owpow_ks	----------	----------	----------	
owrow_cb	----------	----------	----------	
owrow_dw	+s	435	in r	in semivowel
owrow_ks	----------	----------	----------	
owshow_cb	----------	----------	----------	
owshow_dw	+s	91	beg of utt.	within vowel
owshow_ks	----------	----------	----------	
owsow_cb	----------	----------	----------	
owsow_dw	----------	----------	----------	
owsow_ks	----------	----------	----------	
owthow_cb	----------	----------	----------	
owthow_dw	-s	571	end of V2	within vowel
owthow_ks	----------	----------	----------	
owtow_cb	-s	539	end of utt.	within vowel
owtow_dw	----------	----------	----------	
owtow_ks	----------	----------	----------	
owvow_cb	----------	----------	----------	
owvow_dw	----------	----------	----------	
owvow_ks	----------	----------	----------	
owwow_cb	+s	369	release for w	semivowel-vowel
owwow_dw	+s	73	beg of utt.	within vowel
owwow_ks	----------	----------	----------	
owyow_cb	+s	348	release for y	semivowel-vowel
owyow_dw	-s	665	end of V2	within vowel
owyow_ks	+s	323	release for y	semivowel-vowel
owzhow_cb	----------	----------	----------	
owzhow_dw	-s	168	mid V1	within vowel
	-s	745	mid V2	within vowel
owzhow_ks	----------	----------	----------	
owzow_cb	-s	708	end of utt.	within vowel
owzow_dw	----------	----------	----------	
owzow_ks	----------	----------	----------	
uwbuw_cb	----------	----------	----------	
uwbuw_dw	----------	----------	----------	
uwbuw_ks	----------	----------	----------	
uwchuw_cb	-s	100	mid V1	within vowel
uwchuw_dw	----------	----------	----------	
uwchuw_ks	----------	----------	----------	
uwdhuw_cb	----------	----------	----------	
uwdhuw_dw	-s	741	mid V2	within vowel
uwdhuw_ks	-s	593	end of utt.	within vowel
uwdjuw_cb	----------	----------	----------	

Utterance	Landmark	Time	Comment	Hand-classification
uwdjuw_dw	-s	338	closure for dj	vowel-obstruent
uwdjuw_ks	-s	203	closure for dj	vowel-obstruent
uwduw_cb	----------	----------	----------	
uwduw_dw	-s	243	closure for d	vowel-obstruent
uwduw_ks	----------	----------	----------	
uwfuw_cb	----------	----------	----------	
uwfuw_dw	-s	90	beg of V1	within vowel
	-s	582	end of V2	within vowel
uwfuw_ks	----------	----------	----------	
uwguw_cb	----------	----------	----------	
uwguw_dw	----------	----------	----------	
uwguw_ks	+s	53	beg of utt.	within vowel
	-s	603	end of utt.	within vowel
uwhuw_cb	----------	----------	----------	
uwhuw_dw	----------	----------	----------	
uwhuw_ks	----------	----------	----------	
uwkuw_cb	-s	572	end of utt.	within vowel
uwkuw_dw	----------	----------	----------	
uwkuw_ks	----------	----------	----------	
uwluw_cb	+s	377	release for l	lateral-vowel
uwluw_dw	----------	----------	----------	
uwluw_ks	+s	61	beg of V1	within vowel
uwmuw_cb	+s	369	release for m	nasal-vowel
uwmuw_dw	+s	463	release for m	nasal-vowel
uwmuw_ks	+s	305	release for m	nasal-vowel
uwnguw_cb	----------	----------	----------	
uwnguw_dw	----------	----------	----------	
uwnguw_ks	----------	----------	----------	
uwnuw_cb	----------	----------	----------	
uwnuw_dw	+s	482	release for n	nasal-vowel
uwnuw_ks	+s	346	release for n	nasal-vowel
uwpuw_cb	----------	----------	----------	
uwpuw_dw	----------	----------	----------	
uwpuw_ks	----------	----------	----------	
uwruw_cb	+s	336	release for r	semivowel-vowel
uwruw_dw	-s	174	closure for r	vowel-semivowel
uwruw_ks	-s	226	closure for r	vowel-semivowel
uwshuw_cb	----------	----------	----------	
uwshuw_dw	-s	685	end of V2	within vowel
uwshuw_ks	----------	----------	----------	
uwsuw_cb	----------	----------	----------	
uwsuw_dw	----------	----------	----------	
uwsuw_ks	----------	----------	----------	
uwthuw_cb	-s	640	end of utt.	within vowel
uwthuw_dw	----------	----------	----------	

Utterance	Landmark	Time	Comment	Hand-classification
uwthuw_ks	----------	----------	----------	
uwtuw_cb	-s	598	end of V2	within vowel
uwtuw_dw	-s	542	end of V2	within vowel
uwtuw_ks	----------	----------	----------	
uwuw_cb	----------	----------	----------	
uwuw_dw	-s	87	mid V1	within vowel
uwuw_ks	----------	----------	----------	
uwvuw_cb	+s	353	release for v	obstruent-vowel
uwvuw_dw	----------	----------	----------	
uwvuw_ks	----------	----------	----------	
uwwuw_cb	----------	----------	----------	
uwwuw_dw	+s	436	release for w	semivowel-vowel
	-s	608	end of V2	within vowel
uwwuw_ks	+s	354	release for w	semivowel-vowel
uwyuw_cb	-s	174	end of V1	within vowel
uwyuw_dw	+s	295	closure for y	vowel-semivowel
uwyuw_ks	----------	----------	----------	
uwzhuw_cb	----------	----------	----------	
uwzhuw_dw	+s	427	closure for zh	vowel-obstruent
	-s	693	release for zh	obstruent-vowel
uwzhuw_ks	+s	214	closure for zh	vowel-obstruent
uwzuw_cb	+s	214	closure for z	vowel-obstruent
uwzuw_dw	----------	----------	----------	
uwzuw_ks	----------	----------	----------	

Appendix B: Results of the Pivot Analysis

Results of the pivot analysis for the 142 sonorant closures and releases in the VCV database. From this table we determine that pivots capture >96% of all sonorant closures and releases, a significant advantage over sonorant landmarks that examine only >59%.

Utterance	Clos. Or Rel.	Land.	Pivot	Pivot time	+/- g	+/- v	Comment
aalaa_cb	closure	YES	YES	215			
	release	YES	YES	375			
aalaa_dw	closure	NO	YES	283			Not a landmark
	release	NO	YES	477		+v	
aalaa_ks	closure	NO	YES	214			Not a landmark
	release	YES	YES	338			
aamaa_cb	closure	NO	YES	218		-v	
	release	NO	YES	323		+v	
aamaa_dw	closure	YES	YES	274			
	release	YES	YES	427			
aamaa_ks	closure	YES	YES	206			
	release	YES	YES	328			
aanaa_cb	closure	YES	YES	197			
	release	YES	YES	331			
aanaa_dw	closure	YES	YES	304			
	release	YES	YES	471			
aanaa_ks	closure	YES	YES	210			
	release	YES	YES	325			
aangaa_cb	closure	NO	YES	105			Not a landmark
	release	NO	YES	319			Not a landmark
aangaa_dw	closure	YES	YES	255			
	release	YES	YES	452			
aangaa_ks	closure	YES	YES	217			
	release	YES	YES	349			
ahlah_cb	closure	NO	YES	288			Not a landmark
	release	YES	YES	427			
ahlah_dw	closure	NO	YES	248			Not a landmark
	release	NO	YES	473			Not a landmark
ahlah_ks	closure	NO	YES	206			Not a landmark
	release	YES	YES	344			
ahmah_cb	closure	YES	YES	196			
	release	YES	YES	330			
ahmah_dw	closure	NO	YES	232		-v	
	release	YES	YES	447			
ahmah_ks	closure	YES	YES	184			
	release	YES	YES	323			

Utterance	N.Clos. or Rel.	Landmark	Pivot	Pivot time	+/- g	+/- v	Comment
ahnah_cb	closure	YES	YES	169			
	release	YES	YES	329			
ahnah_dw	closure	YES	YES	216			
	release	YES	YES	487			
ahnah_ks	closure	YES	YES	184			
	release	YES	YES	334			
ahngah_cb	closure	YES	YES	169			
	release	YES	YES	359			
ahngah_dw	closure	NO	YES	215			Not a landmark
	release	YES	YES	459			
ahngah_ks	closure	NO	YES	182			Not a landmark
	release	NO	YES	335		+v	
ehleh_cb	closure	NO	YES	220			Not a landmark
	release	YES	YES	368			
ehleh_dw	closure	NO	YES	264			Not a landmark
	release	NO	YES	460			Not a landmark
ehleh_ks	closure	NO	YES	214			Not a landmark
	release	YES	YES	325			
ehmeh_cb	closure	YES	YES	192			
	release	YES	YES	359			
ehmeh_dw	closure	YES	YES	209			
	release	YES	YES	394			
ehmeh_ks	closure	YES	YES	174			
	release	YES	YES	306			
ehneh_cb	closure	YES	YES	147			
	release	YES	YES	345			
ehneh_dw	closure	YES	YES	206			
	release	YES	YES	413			
ehneh_ks	closure	YES	YES	206			
	release	YES	YES	334			
ehngeh_cb	closure	YES	YES	200			
	release	NO	YES	328			Not a landmark
ehngeh_dw	closure	YES	YES	235			
	release	YES	YES	424			
ehngeh_ks	closure	YES	YES	197			
	release	NO	YES	301			Not a landmark
iyliy_cb	closure	YES	YES	190			
	release	YES	YES	327			
iyliy_dw	closure	NO	YES	300			Not a landmark
	release	NO	NO		+g (488)		
iyliy_ks	closure	YES	YES	205			
	release	NO	YES	334		+v	
iymiy_cb	closure	YES	YES	187			

Utterance	N.Clos. or Rel.	Landmark	Pivot	Pivot time	+/- g	+/- v	Comment
	release	YES	YES	294			
iymiy_dw	closure	NO	YES	248			Not a landmark
	release	NO	YES	411			Not a landmark
iymiy_ks	closure	YES	YES	216			
	release	YES	YES	339			
iyniy_cb	closure	YES	YES	190			
	release	YES	YES	326			
iyniy_dw	closure	YES	YES	269			
	release	YES	YES	472			
iyniy_ks	closure	YES	YES	201			
	release	YES	YES	334			
iyngiy_cb	closure	YES	YES	129			
	release	NO	NO		+g (388)		
iyngiy_dw	closure	YES	YES	177			
	release	NO	YES	386			Not a landmark
iyngiy_ks							
owlow_cb	closure	NO	YES	279			Not a landmark
	release	YES	YES	364			
owlow_dw	closure	NO	YES	312			Not a landmark
	release	YES	YES	439			
owlow_ks	closure	NO	NO	-------	-------		Missing pivot
	release	YES	YES	354			
owmow_cb	closure	NO	YES	227			Not a landmark
	release	YES	YES	354			
owmow_dw	closure	NO	YES	316		-v	
	release	YES	YES	475			
owmow_ks	closure	NO	YES	199			Not a landmark
	release	YES	YES	301			
ownow_cb	closure	NO	YES	249			Not a landmark
	release	YES	YES	367			
ownow_dw	closure	YES	YES	311			
	release	NO	YES	477		+v	
ownow_ks	closure	YES	YES	214			
	release	YES	YES	328			
owngow_cb	closure	NO	YES	92			Not a landmark
	release	NO	NO		-g (326)		
owngow_dw	closure	NO	YES	274		-v	
	release	YES	YES	439			
owngow_ks	closure	NO	YES	198			Not a landmark
	release	YES	YES	320			
uwluw_cb	closure	NO	YES	287			Not a landmark
	release	YES	YES	377			
uwluw_dw	closure	NO	YES	364			Not a landmark

Utterance	N.Clos. or Rel.	Landmark	Pivot	Pivot time	+/- g	+/- v	Comment
	release	NO	YES	494			Not a landmark
uwluw_ks	closure	NO	YES	221			Not a landmark
	release	NO	YES	365			Not a landmark
uwmuw_cb	closure	NO	YES	264			Not a landmark
	release	YES	YES	369			
uwmuw_dw	closure	NO	YES	343			Not a landmark
	release	YES	YES	463			
uwmuw_ks	closure	NO	NO	-------	-------		Missing pivot
	release	YES	YES	305			
uwnuw_cb	closure	NO	YES	150			Not a landmark
	release	NO	YES	354			Not a landmark
uwnuw_dw	closure	NO	YES	299			Not a landmark
	release	YES	YES	482			
uwnuw_ks	closure	NO	YES	193			Not a landmark
	release	YES	YES	346			
uwnguw_cb	closure	NO	YES	113			Not a landmark
	release	NO	YES	265			Not a landmark
uwnguw_dw	closure	NO	YES	205			Not a landmark
	release	NO	YES	487			Not a landmark
uwnguw_ks	closure	NO	YES	211			Not a landmark
	release	NO	YES	333			Not a landmark

Appendix C: Design of the FIR filters

In Chapter 4 and 5 we make a reference to two FIR filters designed to measure the energy difference between two frequency bands, the lower being 0-350Hz and higher 350-1000Hz. Filters are designed as FIR because we required a linear phase and constant phase delay across the frequency range. The delay for an N-tap filter as specified in MATLAB is N/2.

In Figure C.1, we illustrate the magnitude and phase response of the 300-tap low-pass filter, designed to pass frequencies up to 350 Hz, designed with the window method and 'hamming' type. Equation used in calculating coefficients for this filter is:

$$B_{low} = fir1(300,380/(sRate/2)) \quad \text{(C.1)}$$

where the sampling rate is 16129 Hz.

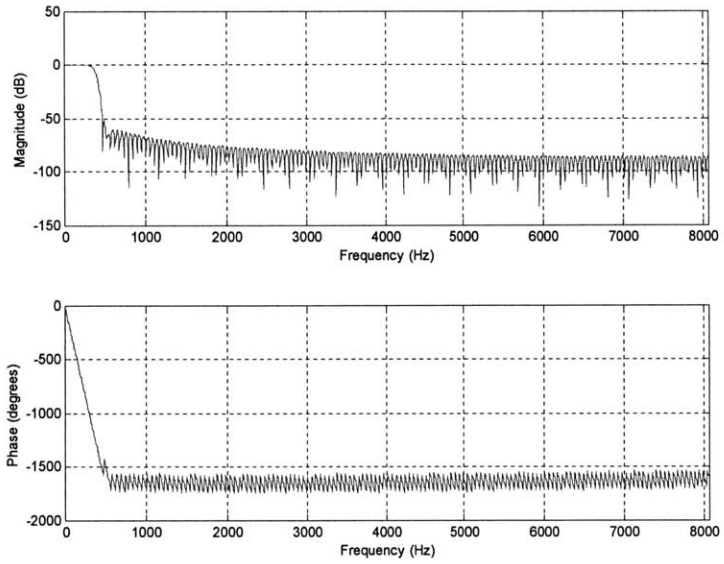


Figure C.1 – Frequency response of the low-pass FIR filter used in the calculation of the cues across the landmark boundary and in nasal murmur.

In Figure C.2 we zoom in to the 0-2000Hz range to show the attenuation at the cutoff rate around 350 Hz.

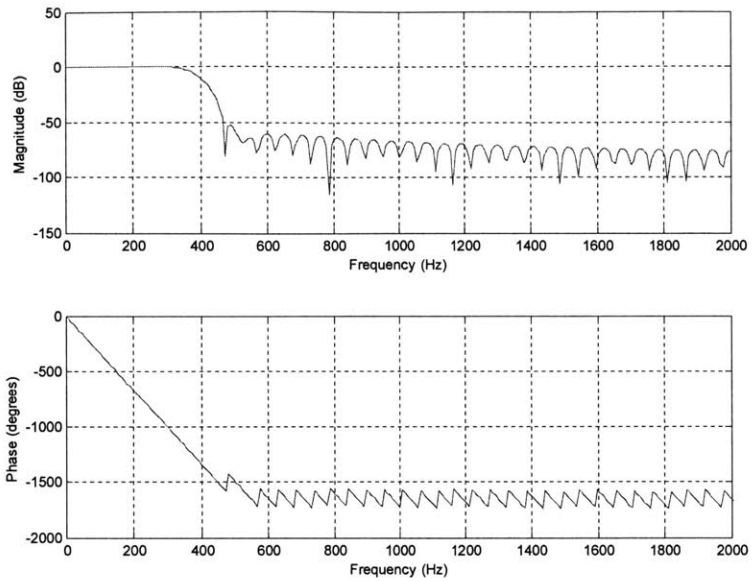


Figure C.2 – Detailed view of the 0-2000Hz frequency range and the attenuation at the cutoff frequency.

In Figure C.3 we illustrate the frequency response of the 300-tap 350-1000 Hz bandpass filter, used in Chapters 4 and 5, and designed with the following equation:

$$B_{bandpass} = fir1(300,[380/(sRate/2),1030/(sRate/2)]) \quad \text{(C.2)}$$

with the sampling rate again at 16129 Hz.

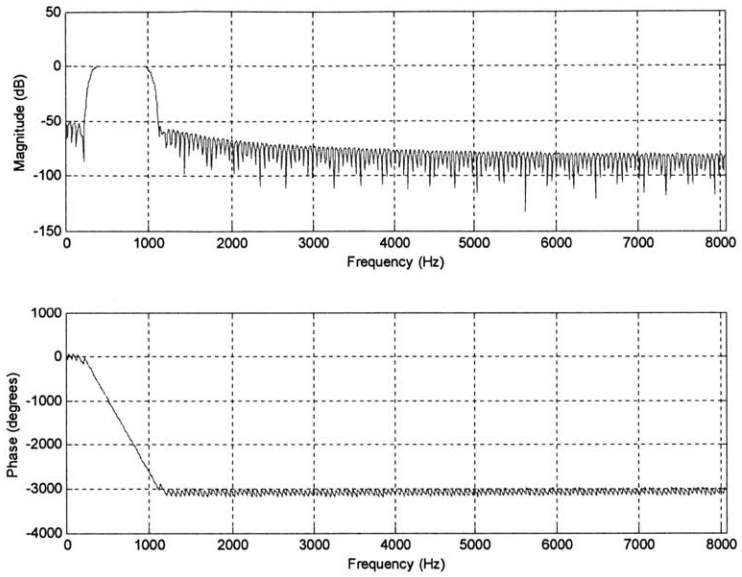


Figure C.4 – Frequency response of the designed 350-1000Hz bandpass FIR filter used in Chapters 4 and 5.

In Figure C.4 we again zoom in on the 2000 Hz range to show the attenuation at the cutoff frequencies.

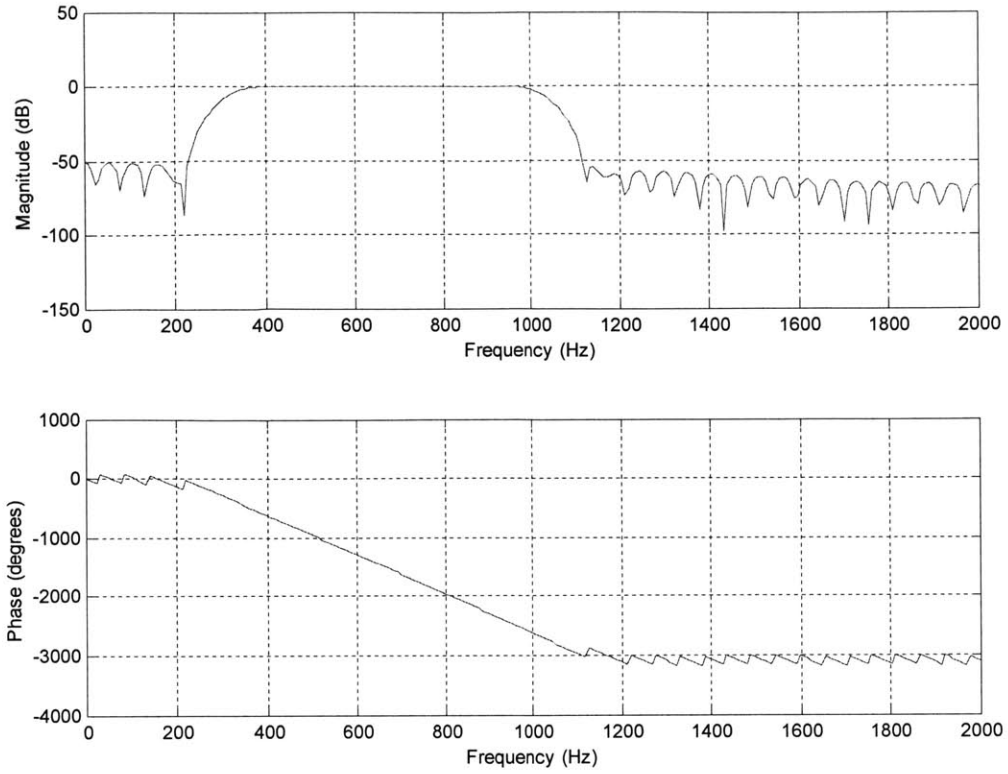


Figure C.4 – Detailed view of the 0-2000Hz frequency range and the attenuation at the cutoff frequency.