Measurement of vowel nasalization by multi-dimensional acoustic analysis

by

Michael A. Berger

Submitted in Partial Fulfillment

of the

Requirements for the Degree

Master of Arts

Supervised by

Professor Joyce McDonough

Department of Linguistics
The College
Arts & Sciences

University of Rochester
Rochester, New York

2007

Curriculum Vitae

Michael A. Berger was born in Niskayuna, New York on March 17, 1970. He attended the University of Massachusetts at Amherst from 1989 to 1992, and graduated with a Bachelors of Arts in Linguistics and Anthropology in 1992. In 1999 he co-founded Speech Graphics Inc., a speech technology company. He came to the University of Rochester in the Spring of 2005 and began graduate studies in Linguistics. He pursued his research in articulatory and acoustic phonetics under the direction of Professor Joyce McDonough and received the Master of Arts degree from the University of Rochester in 2007.

Acknowledgments

I would like to thank my advisor, Professor Joyce McDonough, for all she's done on my behalf: for her steadfast support and encouragement, for opening doors, and for putting me in situations that helped me to grow. This project is an offspring of the "Nasality Group" at Rochester, which includes Joyce McDonough (Department of Linguistics), Meghan Clayards and Neil Bardhan (Department of Brain and Cognitive Sciences). They were integral participants in the planning, hypothesizing and learning that lead to this approach to nasality measurement. Without their collaboration over the past two years this work would never have come into being. They also performed much of the data collection and processing, including Neil's recording of speakers in India. Thanks also to my officemate Jill Thorson, who recorded several speakers on location in Puerto Rico. The Bengali word list used in this study is based on the one provided to us by Professor Aditi Lahiri of the University of Konstanz; Mala Chattergee helped us to understand and transcribe these words: many thanks to both for their valuable assistance. I would also like to thank Professor Michael K. Tanenhaus, who supported this work through an NIH grant ("Time Course of Spoken Word Recognition"), and to whom I am indebted for his unselfish sponsorship and guidance. The presentations in the Tanenhaus lab provided excellent opportunities to gain new insights into this work; I would like to thank the members of that lab for welcoming me into their community. To Richard McGowan of CReSS LLC, I would like to express my gratitude for employing me during my MA program and for helping me learn so much on a variety of topics, and for consulting on aspects of this work. Thanks also to Rebekka, for being so patient and supportive during these busy months. Thanks to my family

for their love and support, and special thanks to my stepparents, who were each integral in their own way to my entering a career in linguistics. Thanks to the entire Linguistics Department, especially Professors Greg Carlson and Jeff Runner, and Daphna Heller, for making me feel at home here; and to Mary Ellen Felten, our superb administrative assistant, for all her help with my program and my thesis. Finally, I would like to thank Professor Christine Gunlogson, without whose friendship I would never have come to this place.

Abstract

When the velum lowers during vowels, the velopharyngeal port is opened, and there is acoustic coupling between the nasal cavity and the main vocal tract, giving rise to a distinct acoustic quality which we call nasality. Nasal coupling results in energy losses at low frequencies, damping of oral formants (especially $F1$), and introduction of nasal formants corresponding to the resonances of the nasal cavity and sinuses. These spectral modifications are gradient, increasing with degree of velar lowering; this relationship suggests that velar position may be recovered from the acoustic signal by measuring the degree of nasality in the vowel. However, the acoustic effects of nasalization vary not only with velar position, but also across different speakers and vowels, making it difficult to isolate an acoustic dimension corresponding to nasalization. This thesis presents a methodology for acoustic measurement of nasality in vowels which attempts to overcome this contextual variability by a normalization procedure. The measuring technique is implemented in an automated measuring system, which is trained on a phonetically balanced set of words recorded from a particular speaker to generate a speaker-specific model. The model specifies the parameters with which to measure nasality for the speaker, the contextual variability of those parameters which may be subtracted for normalization, and the contributions of the parameters to an integrated measure of nasality. The system performs high-frequency nasality measurements over the time course of vowels to generate temporally detailed nasality contours. These measurements may be interpreted as a form of articulatory inference, or as a representation of perceptual input. The system is trained and tested on recordings of 17 native speakers of three languages—English, Spanish and Bengali—speaking three types of vowels: oral (CV(C)), contextually nasalized (C$\tilde{\text{V}}$N), and contrastively nasalized (C$\tilde{\text{V}}$(C)) (Bengali only). The measuring system is evaluated using various performance metrics.

Contents

List of Figures

List of Tables

1 Introduction

In speech acoustics, we attempt to predict the properties of an acoustic signal from a human vocal tract in a particular articulatory configuration. Inversely, we may also attempt to predict the articulatory configuration from the acoustic properties. This is known as the "inverse problem" or "articulatory recovery." This paper confronts the inverse problem in the particular case of recovering velar position from acoustics. More specifically, the focus will be on recovering velar position during vowels.

Nasalization refers to the lowering of the velum during vowels or other oral continuants. When the velum lowers during vowels, the velopharyngeal port is opened, and there is acoustic coupling between the nasal cavity and the main vocal tract, giving rise to a distinct acoustic quality which we call nasality. Vowel nasalization generally occurs as a result of coarticulation between vowels and adjacent nasal consonants: the velar lowering gesture associated with the nasal consonant overlaps with the vowel. Nasal coarticulation happens in both directions—anticipatory and carryover—and can extend across multiple segments and across word or syllable boundaries (see Chafcouloff and Marchal, 1999). Nasal coarticulation is an extremely common event in speech cross-linguistically. In many, but not all, languages there is also a second way vowels can be nasalized: as a contrastive feature. For example, the French words *beau* /bo/ ("beautiful") and *bon* /bõ/ ("good") contrast by nasalization of the vowel.

The acoustic effects of nasalization are well understood. Nasal coupling to the main vocal tract introduces formants and antiformants into the acoustic spectrum. Specifically, it results in energy losses at low frequencies, damping of oral formants (especially $F1$), and introduction of nasal formants corresponding

to the resonances of the nasal cavity and sinuses (Stevens, 1998, pp. 303-322).

Nasalization is gradient: the lower the velum travels, the wider the port opens, and the more nasal the sound. Therefore, velar position should be recoverable from acoustics by somehow measuring the amount of nasality in the signal. If the measure of nasality varies monotonically with velar position, it provides a valid estimation of it—even if the relation is not linear. However, it is difficult to identify such a measure in the acoustics. The reason lies in the general nature of the inverse problem.

1.1 The inverse problem

A half-century of speech science has produced a wealth of knowledge about how the vocal tract produces sounds. The knowledge enables us to predict the properties of acoustic output given a known vocal tract state. However, doing the inverse—recovering the vocal tract state from the acoustic output—is more difficult.

This is in part due to the well-known problem of non-uniqueness in the mapping from acoustics to articulation (e.g., Atal *et al.*, 1978)), by which a single acoustic effect can be caused by several vocal tract shapes. However, even in the absence of a one-to-many mapping, it might still be difficult or impossible to reconstruct the state of the entire vocal tract from an acoustic signal. This is because it requires isolating the effects of individual articulatory factors, which may be conflated in the acoustic output. This is analogous to isolating the effects of individual stones dropped into a pool of water: because of the complex interactions between those effects, it may be impossible to trace back to the original sources.

In the case of nasalization, we seek an acoustic measurement that reflects one articulatory factor: velar position. The problem is that the acoustic effects of velar lowering are conflated with other articulatory factors, including the anatomy of the speaker producing the nasalized sound and the oral articulation (vowel) on which the nasalization is superimposed (Fant, 1960, p. 149).

The nasal cavity, while fixed for a particular speaker, differs widely between speakers. Consequently, for each speaker the acoustic effects of velar lowering will be different. Furthermore, the effect of nasalization varies depending on the shape of the *oral* cavity, which differs not only between speakers, but also between vowels. Consequently, we have a situation where a particular acoustic parameter A is a function of velar position P, as well as speaker S and oral context O.

$$A = f(P, S, O)$$

But what we need in order to infer velar position is a monotonic function of the form

$$A = g(P).$$

Investigators have measured a variety of parameters in the acoustic spectrum in the hopes of finding one or more which, possibly in combination, provide a robust correlate of nasalization. For example, one type of parameter that has been explored in the literature is the amplitude difference between $F1$ and one of the nasal formants (Glass, 1984; Glass and Zue, 1985; Chen, 1995, 1997). In theory, this difference should be monotonic with velar position, because as the velum lowers, $F1$ diminishes in amplitude, while the nasal formant increases. However, due to the dependence on speaker and oral context, such parameters

do not behave in the desired manner.

1.2 Goals

The purpose of this research is to isolate an acoustic measure of nasality which depends on velar position but is largely independent of speaker and oral context. We have strong reasons to believe that this is possible. In many languages of the world, such as Bengali or French, nasalization in vowels is used contrastively. Therefore, members of those speech communities rely on the ability to perceive nasality in order to distinguish words. That is, they must be able to determine when a vowel is more or less nasal. And they must be able to do so regardless of the speaker or the vowel. Since this perceptual ability is based on physical information in the signal, it implies that a "nasality dimension" of some kind, which is strongly related to velar position, must be recoverable in the signal.

This paper presents a novel method for measuring nasality in the acoustic signal. To orient this work, it is important to make the following distinctions regarding methods of nasality measurement:

1. automatic algorithms *vs.* manual procedures
2. quantitative measurement *vs.* classification
3. short-term *vs.* long-term measurement

As to the first distinction, the present research is concerned with developing a rigorous computational algorithm, rather than a procedure that requires manual work such as spectral peak picking. Manual procedures are too slow for large-scale studies, and moreover tend to be less well defined than algorithms that have to be executed by a computer. Regarding the second distinction, unlike

applications in speech recognition that may only require classification of a vowel as nasal or non-nasal, the goal here is a continuous quantitative measure that will reflect the dynamics of velar activity as well as possibly the gradient influence of acoustics on perception. And finally, with respect to the third distinction, we are interested in short-term, high-frequency measurements that will yield a temporally detailed picture of nasality over the time course of a vowel, rather than a long-term value assigned, for example, to an entire vowel or half of a vowel.

Note that this work is concerned with measuring nasality in vowels only, not nasal consonants. Velar lowering leads to very different vocal tract configurations in nasal consonants and vowels. In nasal consonants, output is from the nose only, with the cavity behind the oral closure forming a cul-de-sac resonator. In nasalized vowels, output is from both mouth and nose. Consequently, the acoustic effects of velar lowering in consonants and vowels are different in kind and must be treated separately.

The outline of this paper is as follows. The remainder of this introduction is concerned with the applications of nasality measurement. Section 2 explores the articulatory questions motivating the project. Section 3 reviews previous work in acoustic measurement of nasalization. Sections 4 through 7 present an automated system for measuring nasality over time in vowels. The procedure includes multi-dimensional spectral analysis, followed by a normalization procedure which attempts to remove inter-vowel and inter-speaker variability from the measure. The measure is applied at high temporal resolution over the time course of vowels—specifically at every glottal pulse—which generates detailed temporal profiles of nasality. The measure is developed and tested based on

acoustic data from three languages: English, Spanish and Bengali. Section 8 presents an evaluation of the system and Section 9 gives a general discussion.

1.3 Applications

An automated acoustic measure of nasality over time, if successful, could have a number of applications in research and technology. Possible areas of application include articulatory study, research in human speech perception and automatic speech recognition.

1.3.1 Articulatory research

In this paper, nasality measurement is primarily framed as an articulatory recovery problem; therefore benefits to articulatory studies are naturally emphasized. By recovering velar movement from acoustics we may study the dynamic articulatory processes of nasalization. (See Section 2 for more on the articulatory questions motivating this work.)

But why study articulation through acoustics? One could measure the position of the velum more directly using a variety of devices (see Baken, 1987, ch. 10, for a complete review). These include the nasograph, which measures the degree of velopharyngeal opening by the amount of light passing from a light source in the pharynx to a sensor in the nasal cavity (Ohala, 1971); the velotrace, a mechanical device which rests on the velum and collects analog movement data (Horiguchi and Bell-Berti, 1987); videoendoscopic observation of the velum (Karnell *et al.*, 1988); and EMG measurements of the muscles which control the velum (Ushijima and Hirose, 1974).

Additionally, there are several correlates of velar function which are also

directly measurable. These include nasal air pressure (Weiss, 1954; Shelton *et al.*, 1967); nasal vibration (Stevens *et al.*, 1975, 1976; Horii, 1980); nasal airflow rate (Quigley *et al.*, 1964); and "nasalance," which is the ratio of nasal to nasal+oral acoustic energy output (Fletcher and Frost, 1974). While these measures are not necessarily linearly related to velar position (Amelot, 2004), they seem to be usually monotonically related.

Measuring nasality from the acoustics (that is, from the ordinary oral-nasal output) has several advantages over these other techniques. Unlike all of the measures given above, acoustic measurement is non-invasive, requiring no masks, baffles or probes—merely a microphone and a computer. Hence the subject's normal speech patterns are not impeded. Also, in practice, it should be possible to collect larger amounts of data because there is no need to worry about prolonged discomfort to the speaker. This leads to data sets which are statistically more reliable.

1.3.2 Perception research

A further advantage of an acoustic measurement of nasality is that it simultaneously offers avenues into other areas of research besides articulation—for example, perception. Given that the acoustic signal is the form in which speech actually reaches the listener's ear, measuring nasality in the acoustic medium can be useful in quantifying perceptual input. As illustrated in Figure 1, the acoustic measurement thus serves as a representation for both articulatory reconstruction and perceptual input. Quantification of nasality in the perceptual input could in turn inform studies of nasality perception in speech processing.

There is already a body of literature concerned with quantifying perception

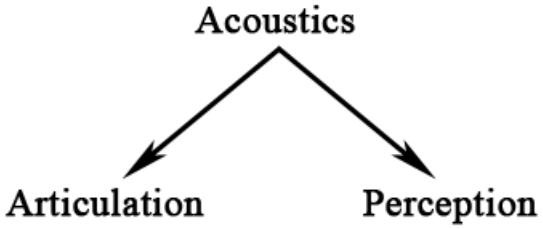

Figure 1: Acoustic measurement as a representation for both articulatory reconstruction and perceptual input.

of nasality in vowels as a function of time. For example, Lahiri and Marlsen-Wilson (1992) (Bengali and English) and Ohala and Ohala (1995) (Hindi and English) used a gating methodology to quantify nasality perception during vowels (to explore the possible effect of lexical representation on speech processing). Research with eyetracking (e.g., Andruski *et al.*, 1994; McMurray *et al.*, 2004) supports the idea that listeners are sensitive to the presence of acoustic cues as they become available in the speech input; the eyetracking paradigm could also be applied to vowel nasalization. However, knowing whether listeners perceive a nasal or non-nasal vowel over time would be more informative if we also knew what information was physically available in the input. A simultaneous *acoustic* measure of nasality could tell us how the information that is available is actually used by listeners: that is, we would have a time-varying measure of both stimulus and response.

1.3.3 Automatic speech recognition

If the information present in the signal is adequate to enable human listeners to distinguish nasal from non-nasal vowels, it should also be sufficient to enable a machine to do the same. Thus an automated acoustic measure of nasalization

in vowels could be of use in automatic speech recognition. It could be of use in detection of phonemically nasal vowels in languages that have them. Furthermore, it could help in detection of nasal consonants, especially in highly coarticulated contexts where the nasal murmur is mostly or entirely absent, and the only indication of the nasal consonant is the nasalization imparted on the vowel (Glass and Zue, 1985; Hasegawa-Johnson *et al.*, 2005). Nasality detection could also be indirectly of use in vowel recognition. Nasalized vowels present a problem for formant tracking algorithms, which become confused by the spectral consequences of nasalization. If nasalization were correctly identified during a vowel, different tracking strategies could be employed (Fant, 1960; Glass and Zue, 1985).

2 Articulation of nasalized vowels

This section briefly explores the articulatory questions motivating this work:

- What is the temporal profile of velar movement in nasalized vowels?
- How does it differ between types of nasalized vowels?
- How does it differ between speakers and languages?

The phrase "temporal profile" denotes the timing, magnitude, duration and speed of velar lowering over the course of the vowel.

In terms of types of nasalized vowels we are concerned primarily with two types: vowels that are nasalized due to coarticulation with a neighboring nasal consonant, and vowels that are nasalized by a contrastive feature associated with the vowel itself. These will be referred to as **contextually** and **contrastively** nasalized vowels, respectively.

This is not to say these are the only causes of velar lowering during vowels. Another cause is **passive velar movement**, in which the velum is displaced due to the movement of other articulators, such as the tongue body. Due to passive velar movement, low vowels tend to be more nasalized than high vowels, even in oral contexts (Moll, 1960). There is also a certain amount of constant **background nasality** that may be present simply because of the speech style or physiology of the speaker. And if there are **structural or functional defects** in the velopharyngeal mechanism, such as cleft palate, the velum may be incapable of closing the velopharyngeal port (Sloan, 2000). **Inadvertent nasalization** is also common among the hearing impaired (Brehm, 1922).

The present research is not centrally concerned with these other forms of nasalization—although it is of interest to ask whether passive velar movement will be reflected in nasality measurements in low *vs.* high vowels. Concerning background nasality, note that the intention is to measure nasality *relative* to this background level.

Studies that compare contextual and contrastive nasalization are rare. Cohn (1990) looked at both contextually and contrastively nasalized vowels in French, and found that in both cases the velum moved more quickly than in English contextually nasalized vowels. She proposed an explanation that in languages such as French, there is pressure to preserve contrast between the two types of vowels, and rapid velar movement helps keep phonemically oral vowels as oral as possible and phonemically nasal vowels as nasal as possible. However, Klopfenstein (2006) looking at another language with contrastive nasalization, Ottawa, found that in contextually nasalized vowels velum movement was not as quick as in contrastively nasal vowels in the same language. Thus the phonemically

oral vowels were not kept as oral as in French. However, in Ottawa, the overall magnitude of nasalization in phonemically nasal vowels was greater than in contextually nasal vowels, suggesting the possibility that in Ottawa the oral/nasal contrast is protected by a different strategy than in French.

In a preliminary study, Berger *et al.* (2007) used the nasality measure presented in this paper to compare velar articulation between three types of vowels—oral, contextually nasal and contrastively nasal—and across three different languages: English, Spanish and Bengali. The results included only average profiles of the vowel tokens of each language, but significant differences between oral and nasal vowels were shown. There was insufficient evidence, however, to suggest a difference between the two types of nasal vowels. A superset of the acoustic data from that study is used here, and similar comparisons of velar behavior over time will be demonstrated below, without averaging over tokens.

There have been various studies comparing the timing and extent of contextual nasalization between languages. Solé (1992, 1995) claimed on the basis of nasograph data that in vowels preceding nasal consonants, the velum is lowered earlier and for a greater portion of the vowel in American English than in as Spanish. Other studies (Clumeck, 1976; Cohn, 1990; Rochet and Rochet, 1991) have found the same difference in timing of nasalization between American English and French (where nasality is contrastive).

3 Acoustic measurement of nasalization

House and Stevens (1956) constructed an idealized electrical analog of the vocal tract to study the acoustic effects of gradually opening the velopharyngeal port during vowels. They concluded that the major effects of nasalization were the

reduction in amplitude of the first formant, with concomitant broadening of its bandwidth; upward shifting in the frequency of the first formant; and an overall reduction in the energy of the vowel. Additionally, the emergence of a spectral prominence above $F1$ at around 1000 Hz was observed. Hattori *et al.* (1958) found another nasal resonance below $F1$ between 250 and 450 Hz. These effects are illustrated in Figures 2 and 3. Fant (1960) confirmed these general characteristics of nasalized vowels, but noted that the exact acoustic consequences of nasalization vary considerably between different vowels and speakers.

Researchers attempting to measure nasality in vowels have tried to reduce these acoustic modifications to one or several key parameters. Most efforts have involved looking for acoustic parameters that robustly correlate with *perception* of nasality when manipulated in synthetic speech. Studies attempting to establish a correlation with velar position (by comparing acoustic parameters to articulatory data) are relatively rare.

House and Stevens (1956) manipulated the amplitude of $F1$ ($A1$) in their analog synthesizer, and had listeners judge whether the sound they heard was nasal or non-nasal, or give a judgment of degree of nasality. They found that $A1$ needed to be reduced by 8 dB for positive nasality judgments to reach 50%. Huffman (1990) looked at changes in the relative amplitude of $F1$ instead of absolute amplitude, by taking the difference between $A1$ and $H1$ (the amplitude of the first harmonic). Measurements of $A1 - H1$ in natural speech, both averaged over the vowel and changing over time, were correlated with listeners perception of nasal *vs.* oral vowels. The study found that both the average value of $A1 - H1$ and the direction of change of the parameter over the time course of the vowel contributed to perception of nasality. These results demonstrated

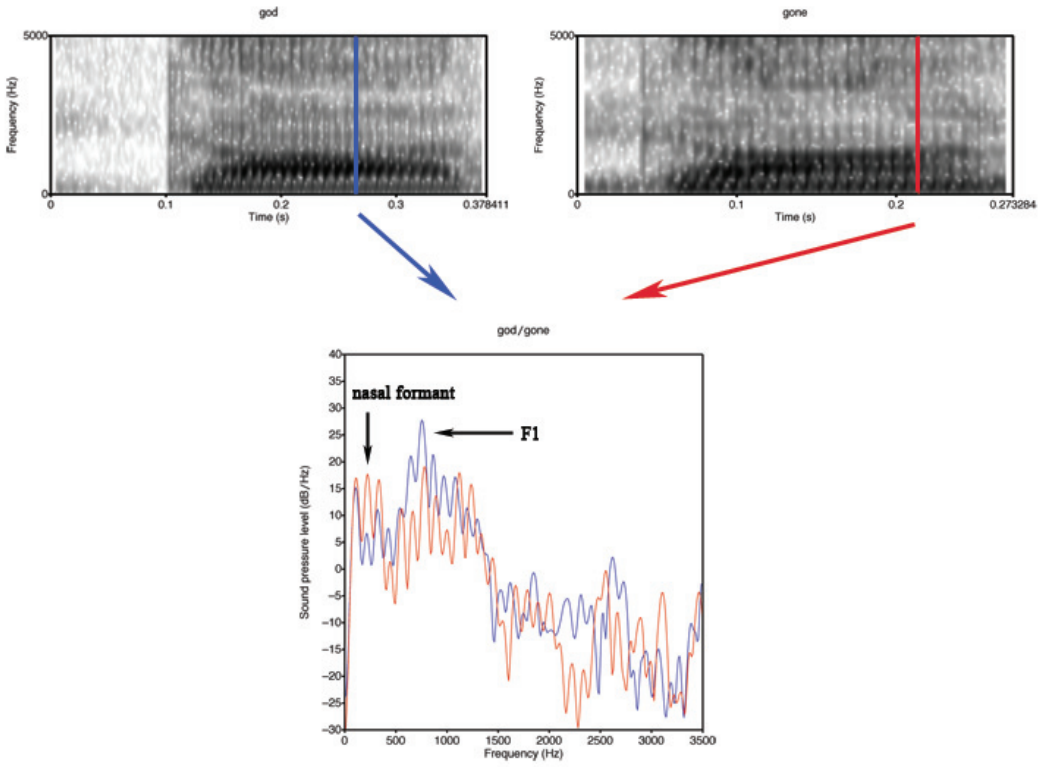

Figure 2: Spectral comparison of oral and nasalized /a/. The top two panels show spectrograms of the words *god* (oral /a/) and *gone* (contextually nasalized /a/) spoken by a male English speaker (E16). The bottom panel compares spectra taken from the oral vowel (blue) and nasalized vowel (red) at the time points indicated by vertical lines. The comparison highlights the flattening of $F1$ in the nasal vowel, as well as the emergence of the nasal formant below $F1$.

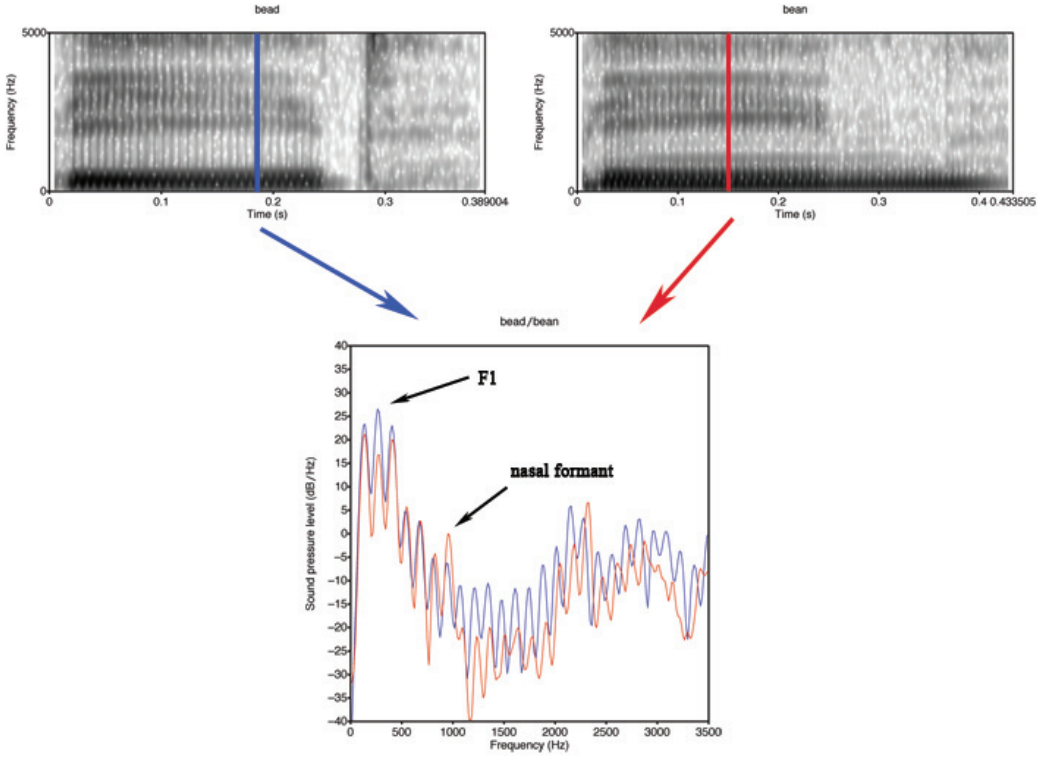

Figure 3: Spectral comparison of oral and nasalized /i/. The top two panels show spectrograms of the words *bead* (oral /i/) and *bean* (contextually nasalized /i/) spoken by a male English speaker (E13). The bottom panel compares spectra taken from the oral vowel (blue) and nasalized vowel (red) at the time points indicated by vertical lines. The comparison demonstrates amplitude reduction and a slight frequency increase in $F1$ in the nasalized vowel, along with the introduction of a nasal formant above $F1$ and a general reduction in energy.

the importance of dynamic measurement to perceptual studies.

The introduction in synthetic speech of the nasal peak above $F1$ (around 1000 Hz) (Maeda, 1982; Hawkins and Stevens, 1985) or the other nasal peak below $F1$ (around 250-450 Hz) (Hattori *et al.*, 1958; Maeda, 1982) was also found to enhance nasality perception for particular vowels. Chen (1995, 1997) proposed two parameters which combined relative $A1$ decrease with the emergence of the two nasal formants: these were $A1-P0$ and $A1-P1$, where $P0$ is the amplitude of the nasal formant below $F1$, and $P1$ is the amplitude of the nasal formant above $F1$. As nasality increases, $F1$ should decrease while $P0$ and $P1$ increase; thus both $A1-P0$ and $A1-P1$ should decrease with increasing nasal coupling. In synthetic speech, Chen found a higher correlation of these parameters to perceived nasality than $A1$ alone.

Chen also proposed a modification of these two parameters to attempt to make them independent of vowel context. Acknowledging that the proximity of $F1$ or $F2$ to one of the nasal formants could add a boost to $P0$ or $P1$ that is not due to nasality, she attempted to subtract out this influence using a normalization formula based on the frequencies and bandwidths of the nasal formants.

Chen's approach to measuring nasalization in vowels has several advantages: it uses relative rather than absolute measurements of amplitude; it incorporates multiple effects of nasalization (reduction of $A1$ combined with increases of $P0$ and $P1$) rather than relying on a single measure; and it attempts to normalize these parameters over vowel types. However, there are also several problems with the approach.

First, the two measures do not permit uniform measurement for all vowels,

since different vowel types are more conducive to one measure or the other. This is because the oral formants can occlude the nasal formants, making them inaccessible to measurement: the low nasal formant may be hidden when $F1$ is low (as in high vowels), and the high nasal formant may be hidden by either a high $F1$ (low vowels) or low $F2$ (back vowels). Figure 4 shows two spectra in which $P0$ and $P1$ are each obscured by $F1$.

Second, tracking nasal formants is even more difficult than tracking vowel formants. Nasal formants emerge at different frequencies for different speakers and also drift within the speech of one speaker. They are also much less salient for some speakers than for others. Tracking nasal formants is particularly difficult with female speakers, for two reasons: resonances in general are broader and less prominent in vowels spoken by females; and due to higher fundamental frequency, the harmonics are more widely spaced in the frequency domain, increasing the chance that the nasal resonances will disappear in the gaps between harmonics. Figure 5 illustrates these differences in formant prominence and harmonic resolution between spectra of a male and female speaker. (The general spectral differences between male and female speakers is in fact a prime example of how the acoustic effects of nasalization will depend on speaker anatomy.)

Note finally that Chen's measurement technique was not automated—it involved manual identification of the relevant peaks in each spectrum (which seems to be required given the problems noted above). Also, measurements were long-term averages (taken at three points in the vowel) rather than high-frequency, short-term measurements that would generate a temporally detailed profile of the vowel.

In literature relating to speech recognition, there are various studies con-

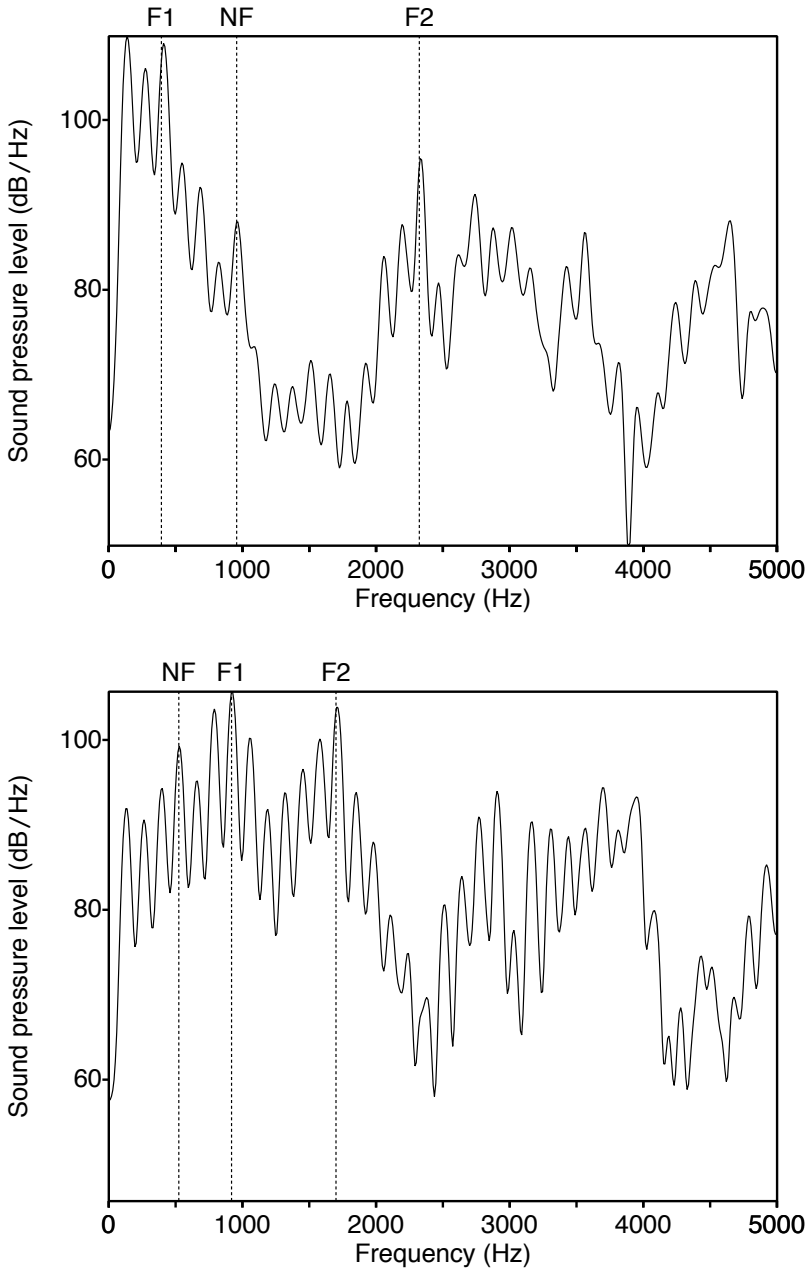

Figure 4: Two spectra in which nasal formants are obscured by oral formants. The top spectrum is a sample from the high vowel [i] in the word "bean" spoken by a male English speaker; the low $F1$ obscures the nasal formant that is predicted in the 250-450 Hz region (but note the unobscured nasal formant at 1000 Hz (NF)). The bottom spectrum is from the low vowel [a] in the word "santo" spoken by a male Spanish speaker; the high $F1$ obscures nasal formant that is predicated in the vicinity of 1000 Hz. (The peak at 500 Hz may be the unobscured low nasal resonance.)

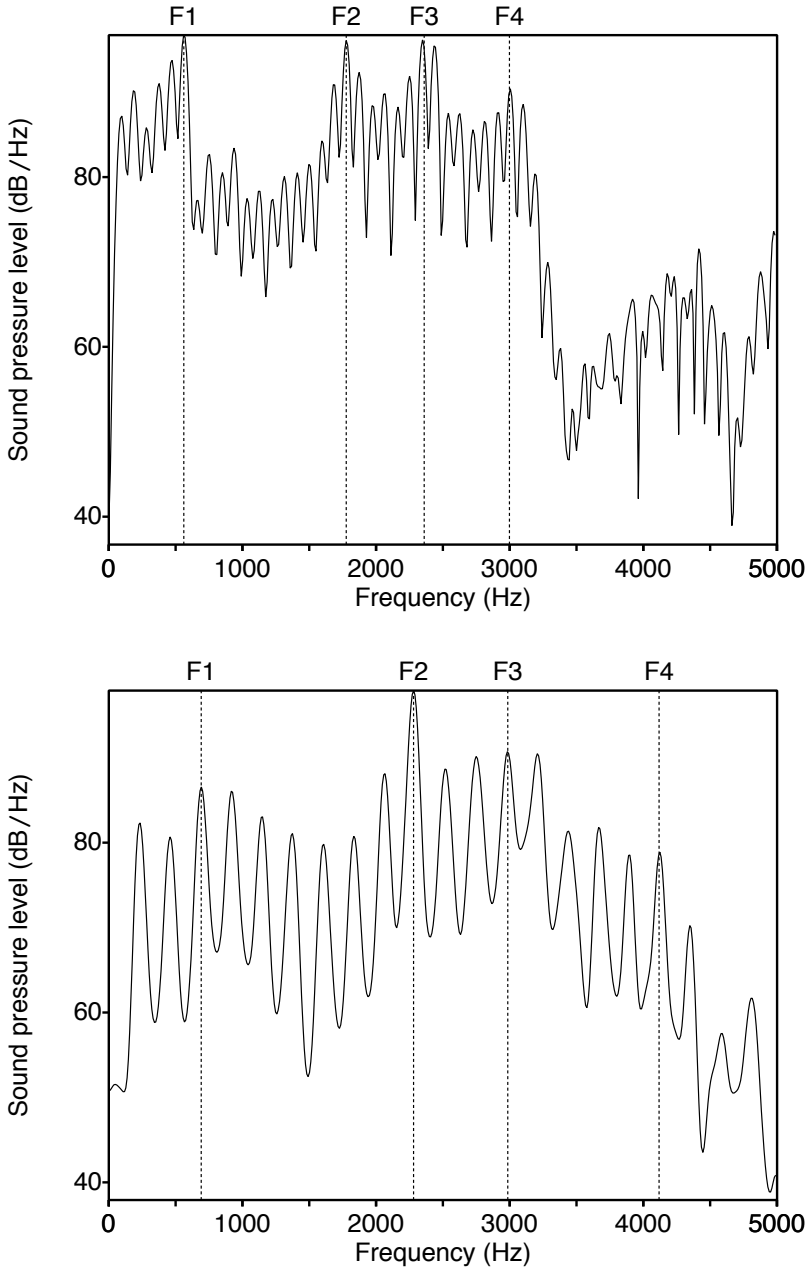

Figure 5: Spectral frames taken from the midpoint of the vowel /æ/ in *pan* spoken by a male (top) and female (bottom) speaker. Peaks of the first four formants are indicated. The female speaker has less prominent resonances and poorer harmonic resolution, making formant tracking more difficult.

cerned with detection of nasalized vowels. Note that such studies are concerned with classifying a vowel as nasal or non-nasal (often to help determine whether the following consonant is nasal), rather than with giving a gradient measurement of nasality. For example, Glass (1984) and Glass and Zue (1985) used a set of six acoustic parameters to automatically detect whether a vowel is nasalized. These were (1) the center of mass below 1000 Hz, (2) the standard deviation around the center of mass, (3) the minimum percentage of the time there is a nasal resonance in the low-frequency region, (4) the maximum percentage of time there is a nasal resonance in the low-frequency region, (5) the maximum value of the average amplitude dip between $F1$ and the nasal resonance, and (6) the minimum value of the average amplitude difference between $F1$ and the nasal resonance. The parameters were measured in each of three subregions of the vowel. Using a sum of the individual log likelihoods calculated from the parameters, they were able to achieve a correct detection rate of 74% in a corpus of 200 words recorded from six speakers. They also found that the system performed significantly better for males than for females.

More recently, Pruthi (2007) evaluated 37 acoustic parameters and cited nine of these as best-performing in an automated nasality detection task. Using the nine parameters he achieved accurate nasality detection rates of 96%, 78% and 70% for StoryDB, TIMIT and WS96/97 corpora, respectively.

4 System components

The nasality measurement technique presented in this thesis is based on the studies of acoustic parameters in the literature. However, it attempts to process these parameters in such a way as to find a correlation to velar position that is

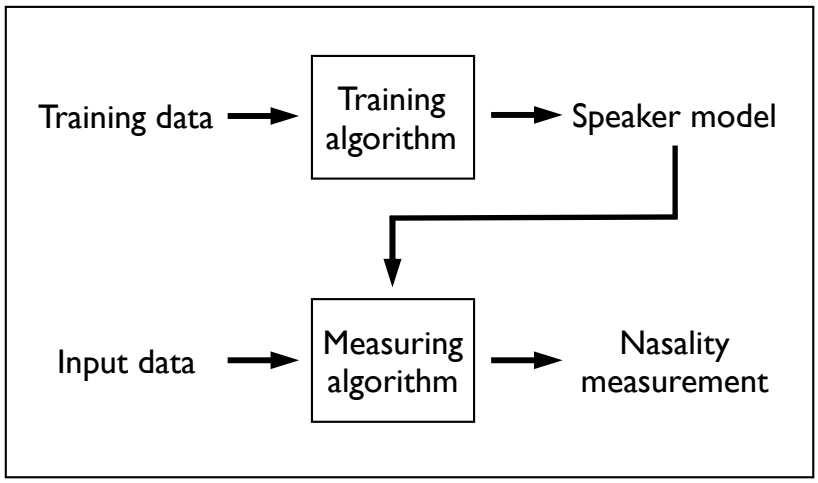

Figure 6: Functional arrangement of the measuring system.

independent of speaker and vowel. In addition, unlike most previous methods, the measuring technique used here is implemented computationally, in a fully automated software system. Furthermore, this system is designed to produce quantitative measurements of nasality rather than merely binary (oral/nasal) classification; and short-term, high-frequency measurements rather than measurements of the vowel as a whole or of large parts of it.

The diagram in Figure 6 shows the functional arrangement of the measuring system reported here. The measuring system comprises a *training algorithm*, which derives a *speaker model* from acoustic training data, and a *measuring algorithm*, which uses the speaker model to perform nasality measurements on acoustic input data. Training is speaker-specific: a separate model is generated for each speaker, based on the statistical distributions of various acoustic parameters, and that model is used to perform nasality measurements on input recordings from the same speaker. The algorithms have been implemented in a set of programs written in the Praat (Boersma and Weenink, 2007) scripting language and MATLAB. The following sections describe these components in more

detail. Section 5 describes the training data; Section 6 the training algorithm and resulting speaker model; and Section 7 the measuring algorithm.

5 Training data

The measuring system is trained on a set of annotated digital audio recordings of each speaker. The acoustic data used in this study was gathered and processed with the help of colleagues at the University of Rochester: Meghan Clayards and Neil Bardhan in the Department of Brain and Cognitive Sciences, and Joyce McDonough and Jill Thorson in the Department of Linguistics. We recorded native speakers of three languages—English, Spanish and Bengali—reading from prepared word lists. All three languages exhibit contextual nasalization of vowels, and Bengali also has contrastive nasalization.

5.1 Word lists

The word list for each language was constructed out of "minimal sets" of lexical items designed to exemplify differences between the three classes of vowels under consideration: oral, contextually nasal and contrastively nasal. In all three languages, the minimal sets included minimal pairs contrasting oral and contextually nasal vowels; and in Bengali only, there were also minimal triples, which further included contrastively nasal vowels. More specifically, minimal pairs were generally of the form {CV(C), C$\tilde{\text{V}}$N}, and minimal triples were of the form {CV(C), C$\tilde{\text{V}}$N, C$\tilde{\text{V}}$(C)}—where each C is an oral consonant (or cluster of oral consonants), each N is a nasal consonant (or in rare instances a nasal consonant followed by a homorganic oral stop), and V and $\tilde{\text{V}}$ represent the same vowel in oral and nasalized forms. In a few cases the words were not monosyllabic but

	English	Spanish	Bengali
CV(C)	/hæg/ "hag"	/dos/ "two"	/ɖɑl/ "lentils"
C$\tilde{\text{V}}$N	/hæŋ/ "hang"	/don/ "sir"	/ɖɑn/ "right"
C$\tilde{\text{V}}$(C)			/ɖɑ̃ʈʰ/ "arrogant"

Table 1: Examples of English and Spanish minimal pairs and a Bengali minimal triple.

the target syllable always matched one of the patterns above.

Table 1 gives examples of minimal pairs in English and Spanish, and a minimal triple in Bengali. In the ideal case, the words in a minimal set differed only in one phonological feature: either the nasality of the coda consonant or that of the vowel. In reality, due to constraints on lexical inventories, we had to allow differences in other features, such as place or manner of articulation of a consonant in the context. However, consonants in the syllable onset were strictly required to be non-nasal.

In selecting lexical items, preference was given to minimal sets in which the oral consonants were obstruents rather than sonorants. This is because sonorants, being more vowel-like, place greater demands on the articulators used for vowels and hence are expected to have stronger coarticulatory influence on vowels. It was desired to minimize this influence so as to maximize the homogeneity of each vowel. Furthermore, unlike obstruents, sonorants lack salient landmarks that make it clear where to place segment boundaries, such as the onset or release of a constriction.

The strategy of constructing the word lists from minimal sets contrasting oral and nasal vowels led to the result that the word lists were fairly balanced between oral and nasal vowels. To the extent that the lexicon allowed, we also sought to give equal representation to each of the vowel qualities in the inventory. Thus both nasal category and vowel quality were variables that were fairly well

balanced in the data. Moreover, since vowel quality was constant within each minimal set, a further property was that within each vowel quality there was a balance between nasal categories, and vice-versa. This phonetic balancing of the word lists turns out to be very helpful for the statistical analysis described below.

Working under these criteria we produced an English word list with 90 items and a Spanish word list with 34 items. A Bengali word list largely fitting our criteria was provided to us by Dr. Aditi Lahiri of the University of Konstanz, and after some minor changes that list contained 109 words.[1] The three word lists are provided in Appendix A.

5.2 Data acquisition

We recorded four speakers of Bengali, six speakers of American English, and seven speakers of Spanish. The 17 speakers include 10 males and seven females, ranging in age from 17 to 69. All are native speakers of their respective languages. Some of the speakers were recorded in the field in Calcutta, India and Caguas, Puerto Rico. Recording was performed using a Marantz PMD 670 digital recorder and a unidirectional microphone at 44.1 kHz. Each speaker's recording was done in a single session.

Most of the speakers produced the words in a carrier phrase rather than in citation form. Carrier phrases were used by all seven Spanish speakers, four of the six English speakers, and two of the four Bengali speakers. The carrier phrases were as follows: in Spanish, "Di ____ fuerte" (*Say ____ loudly*); in English, "Say ____ again"; and in Bengali, "Abar ____ bolo" (*Again ____ say*).

[1] Any errors in the current Bengali word list are no doubt our own.

ID	Sex	Age	Carrier Phrase
B11	F	50s	no
B12	M	61	no
B13	M	17	yes
B14	M	19	yes
E11	F	25	yes
E12	F	25	yes
E13	M	20s	no
E14	M	36	no
E15	M	59	yes
E16	F	50s	yes
S11	M	20s	yes
S12	M	20s	yes
S13	F	22	yes
S14	F	58	yes
S15	M	20	yes
S16	F	69	yes
S17	M	13	yes

Table 2: Speakers

One of the Bengali speakers (B14) did not use this carrier phrase but rather produced each word in a different sentence he composed. Most speakers repeated each item (with carrier phrase) three times in a row. The speakers are listed in Table 2 with sex, age and carrier phrase information. (The first letter of each speaker's ID indicates the language spoken.)

5.3 Annotation

The input to the training algorithm consists of a set of sound files. Within each sound file, the acoustic analysis targets the intervals corresponding to the vowels in the word lists. The sound files must be manually annotated in order to make those vowel intervals accessible to the algorithm; that is, the intervals must be defined in the time domain and labeled. We used Praat TextGrids (Boersma

and Weenink, 2007) to do this annotation.

The defining of discrete intervals ("segmentation") on the continuous speech stream is more of an art than a science. Nonetheless, the placement of vowel boundaries is of some significance because it defines the acoustic material on which the algorithm operates. The strategy for segmenting vowels was roughly as follows. If the vowel was preceded by an obstruent, the dividing boundary was generally placed at the first glottal pulse after the closure release. (This meant aspiration after stop releases was excluded from the vowel.) Similarly, if the vowel was followed by an obstruent (or a nasal stop), the boundary was placed at the last glottal pulse before the onset of closure. For the less straightforward case of dividing the vowel from an adjacent oral sonorant, we sought the balance-point between the two sounds based on auditory judgments and visible formant transitions.

6 Training algorithm

This section presents the training algorithm, which derives a set of speaker-specific information, called a speaker model, from the acoustic training data (described in the preceding section). The model specifies the parameters with which to measure nasality for the speaker, the contextual variability of those parameters which may be subtracted for normalization, and the contributions of the parameters to an integrated measure of nasality. More specifically, a speaker model comprises these three components:

1. A set of *nasality parameters* $P = \{p_1, p_2, ..., p_n\}$. These are the acoustic parameters through which nasality will be measured; each of them is

expected to correlate with velar position.

2. A set of *context-dependent parameter distributions* $\{D_p\}, p \in P$. These are normal distributions of the nasality parameters dependent on oral context (the configuration of the oral cavity). The distribution of each parameter p is expressed as a pair of functions $D_p = (\mu_p(o), \sigma_p(o))$, which give the mean and standard deviation of the parameter as a function of oral context o.

3. A *parameter integration function* $I(\vec{x})$. This is a function applied to vectors $\vec{x}$ of nasality parameter values to reduce them to scalar values constituting measures of nasality.

Training is speaker-specific: a separate model is generated for each speaker. Once established, the speaker model may be used by the measuring algorithm to measure nasality in vowels recorded from the same speaker.

The training algorithm comprises three main steps: (1) discretization of the input data by sampling and parameterization; (2) acquisition of the context-dependent parameter distributions; and (3) derivation of the parameter integration function. These steps are discussed in sequence in the following subsections.

6.1 Sampling and parameterization

The first stage of the training algorithm is to reduce the acoustic training data to a discrete form. The waveforms are *temporally* discretized by sampling each vowel interval at specified points in time. Specifically, the vowels are sampled at the extremum of each glottal pulse. The effect of sampling at glottal pulse extrema is to phase-align the samples; that is, each sample will be located at

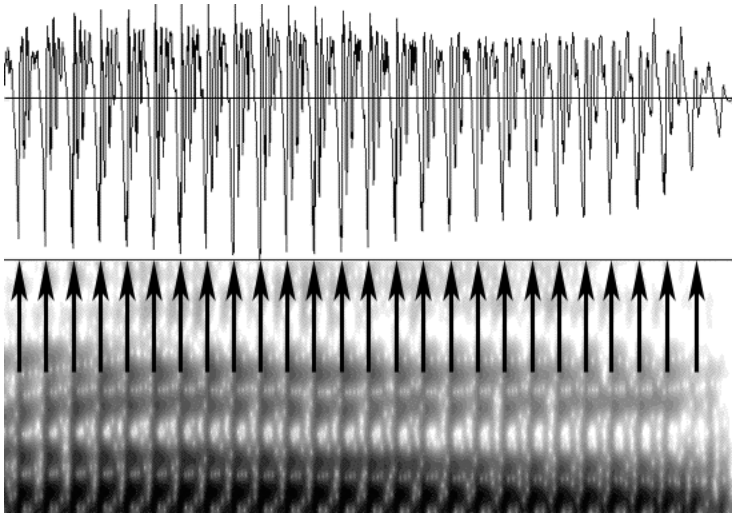

Figure 7: Sampling of a vowel at glottal pulse extrema. The top panel shows the waveform and the bottom panel the spectrogram.

the same phase of the glottal pulse cycle. Phase-aligning the samples seems to minimize differences (jitter) between short-time spectra centered at each sample (Richard Aslin, personal communication). Note that a consequence of sampling once per glottal pulse is that the frequency of samples is tied to the pitch of the speaker. As a result, females tend to yield more samples than males for the same speech task. (Interestingly, this bias may be counterbalanced by the lower harmonic resolution in females' spectra.) Glottal pulse peaks are identified using Praat's pitch tracker and pulse-finding algorithm. Figure 7 illustrates the sampling of a vowel.

Each sample is converted into a finite vector of values by parameterization of the acoustic signal at that time point. The parameters measured include the first and second formant frequencies, and the nasality parameters, which as discussed above are a set of acoustic parameters expected to correlate with velar position. $F1$ and $F2$ are measured using Praat's formant tracker (Burg method)

set to seek five formants below 5000 Hz for males and 5500 Hz for females, using a 25-millisecond window.

The choice of nasality parameters to use is a variable in the speaker model. One of the objectives here is to develop methods of evaluating acoustic parameters to assess their eligibility as nasality parameters. To keep the discussion simple, only the following candidates will be considered: $A1-H1$, COG (center of gravity in the low frequency spectrum), $B1$ (bandwidth of the first formant), $A1-P0$ and $A1-P1$. As noted above, $A1-H1$, or relative $F1$ amplitude, should decrease with nasalization due to the flattening out of $F1$; for the same reason, $B1$ should increase. $A1-P0$ and $A1-P1$ should decrease as described by Chen. Center of gravity, or center of mass, is a mean of frequencies in the spectrum weighted by amplitude. Center of gravity in the low end of the spectrum—up to 1000 Hz—is expected to drop with increasing nasalization due to addition of a nasal formant in the vicinity of $F1$ (Glass, 1984). For the sake of comparison, center of gravity up to 1500 Hz will also be tried.

These parameters (or their components) are measured as follows. Values for $B1$ are obtained through Praat's standard bandwidth command, which defines the bandwidth of a formant as the width of the formant in the LPC-smoothed spectrum at 3 dB below the peak. The remaining values are obtained on the basis of a narrow-band spectral analysis performed over a 30-millisecond Gaussian window centered at the sample time.

$A1$ (peak amplitude of $F1$), $H1$ (amplitude of the first harmonic), $P0$ (amplitude of the nasal formant below $F1$) and $P1$ (amplitude of the nasal formant above $F1$) are all amplitudes of selected harmonics in the spectrum; so first, all of the harmonics must be located using a harmonic-finding algorithm. The

harmonic-finding algorithm finds all of the maxima in the spectrum, defines the first harmonic as the one closest to the fundamental frequency returned by Praat's pitch function, and recursively defines each subsequent harmonic as the maximum whose offset from the previously found harmonic is closest to one increment of the fundamental frequency.

Once the harmonics are located, $H1$, $A1$, $P0$ and $P1$ are measured as follows. $H1$ is simply the amplitude of the first harmonic. $A1$, $P0$ and $P1$ are formant peaks measured by finding the most prominent harmonic in a certain frequency region. (It may be easier to find the formant peaks if the signal is first pre-emphasized to control for spectral tilt, though this was not done in the present study.) To measure $A1$, the algorithm searches for the highest-amplitude harmonic within 1.2 increments of the subject's fundamental frequency at the sample time. Given the aforementioned difficulties of tracking nasal formants (which were encountered first-hand in an earlier version of the algorithm), $P0$ and $P1$ are not obtained by any attempt at formant tracking, but simply by taking the amplitude of the most prominent harmonic in the ranges where the nasal formants are expected to arise. The ranges used were 0 to 450 Hz for $P0$, and 800 to 1100 Hz for $P1$. While this is an admittedly crude approximation of these variables, it is at least a straightforward way to use them in an automated process.

To select the nasality parameters, the candidates will be comparatively evaluated using two criteria: discrimination and average acceleration. These criteria will be defined in Section 6.1. In a future stage of development, a superset of nasality parameters may be automatically ranked using these criteria to determine which are the most reliable indicators of nasality for a particular speaker.

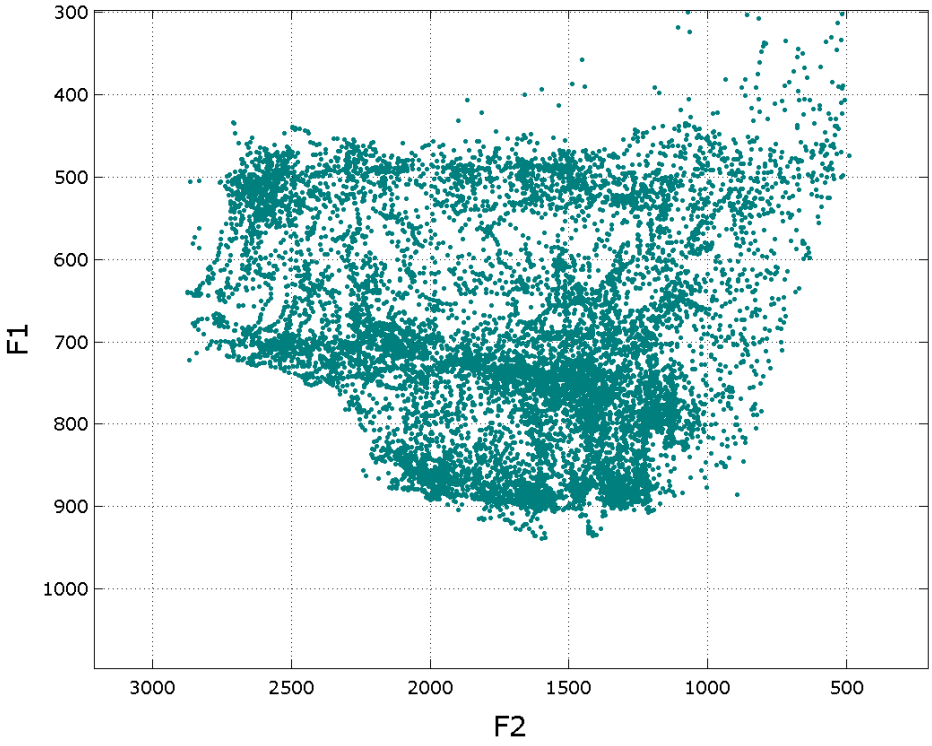

Figure 8: Plot of speaker E11's entire sample set in formant space. This sample set contains 11,164 samples.

The ranking could then be used to select which parameters to include in the speaker model, or to determine a weighting of the parameters in the parameter integration function. This selection or weighting process would be an additional step in the training algorithm itself.

The sampling and parameterization of the acoustic data results in a large number of observations in acoustic parameter space. To illustrate, Figure 8 shows all of the samples from speaker E11 plotted in formant space: a total of 11,164 samples. Of course, the number of samples depends on several factors, including the number of items in the word list, the number of repetitions of each

word, the lengths of the vowel tokens produced by the speaker, and the pitch of the speaker's voice (which determines sampling frequency).

6.2 Acquiring the nasality parameter distributions

The nasality parameters vary not only as a function of nasal coupling, but also due to differences in oral configuration and speaker anatomy. The goal is to obtain a nasality measure that is largely independent of these other articulatory factors. Rather than seeking an acoustic parameter that somehow transcends these influences—i.e., a robust global measure—the approach taken here is to characterize the variability of the nasality parameters over speaker and oral context, and remove that variability by a normalization process.

The contextual variability of the nasality parameters is characterized by determining the distributions of the parameters in each local context. For this purpose, the speaker model will contain a set of context-dependent parameter distributions. Each distribution in the speaker model, D_p, is a pair $(\mu_p(o), \sigma_p(o))$ giving the mean and standard deviation of the parameter p as a function of oral context o for that speaker. As described below, each parameter is normalized across oral contexts using z-score transformations based on these local distributions.

The context-dependent parameter distributions are inferred from the sample set of the speaker. In order to determine the mean and standard deviation of a parameter in a particular oral context, it is necessary to define the oral contexts and map the samples to them. Two alternative conceptions of oral context are explored here. In one, the oral context of a particular sample is the phoneme class of the vowel token from which the sample was taken. In another, the oral

context is defined as the position of the acoustic sample in formant space. The acquisition of parameter distributions dependent on both types of oral context are described below.

6.2.1 Distributions dependent on vowel type

Vowel phonemes provide a simple characterization of the oral context of an acoustic sample. This characterization is based on the assumption that a phoneme represents a stable articulatory configuration in the oral cavity. The phoneme affiliation of a given acoustic sample is easy to obtain. During annotation (see Section 5.1), appropriate phoneme labels were associated with each vowel token in the recording. (These labels were based on a "broad" transcription of the vowels and no attempt was made to represent free variation or inter-speaker differences in pronunciation.) The phoneme affiliation of a sample may be determined from the label of the vowel token from which it was taken. Note that for purposes of defining oral contexts, phonemically nasal vowels in Bengali are classed together with their oral counterparts.

Once samples are mapped to phonemes, context-dependent parameter distributions can be obtained by taking the mean and standard deviation of each parameter within each phoneme's sample population. For example, Figure 9 illustrates the means and standard deviations of the parameter $A1 - H1$ within each of the English vowel phonemes for speaker E14. The influence of oral context on the nasality parameter is evident from the differences between the local distributions.

We now introduce the normalization function that converts parameter measurements to a standard distribution. Within the speaker model, $\mu_p(v)$ and $\sigma_p(v)$

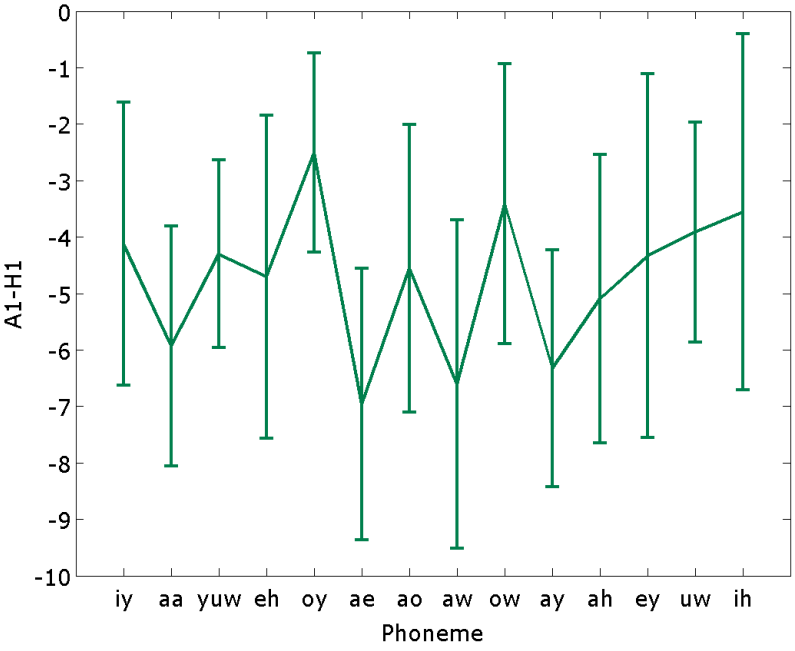

Figure 9: Means and standard deviations of $A1 - H1$ by phoneme based on the sample set of speaker E14 (English-speaking male).

are the mean and standard deviation, respectively, of a nasality parameter p in the sample population of a vowel phoneme v. For any sample affiliated with phoneme v, its original value in p, x, is adjusted using the z-score normalization function

$$N(p, x, v) = \frac{x - \mu_p(v)}{\sigma_p(v)} \tag{1}$$

Note that since the normalization function is a z-score, each parameter after normalization will be centered around zero and given in units of standard deviation.

It is important to note that the success of this normalization procedure depends on the accuracy of the parameter distributions within each oral context. The accuracy of the distributions in turn depends on the phonetic balancing of the training data (Section 4.1). A balance between vowel phonemes helps secure a large enough sample population for each phoneme to infer a distribution. Moreover, a balance between nasal and non-nasal tokens within each phoneme promotes the result that the mean will represent a true midpoint in nasalization, and the standard deviation will reflect the true range of cases.

As discussed in Section 7 below, the normalization function is used by the measuring algorithm to normalize parameter measurements taken from acoustic data of the same speaker. That acoustic data is not required to be phonetically balanced, but the vowel intervals must still have phoneme labels, so that each sample's oral context can be identified for normalization of the sample.

6.2.2 Distributions dependent on position in formant space

Vowel phonemes provide a convenient way to divide a speaker's sample set into groupings reflecting local oral configurations. However, there are a number of

drawbacks to dividing the sample set in this way.

First, there is the obvious drawback of being required to manually label the vowel tokens in the acoustic input, both for training and for measurement. Second, and more importantly, this approach assumes that vowel tokens with the same phoneme label are articulatorily similar, although this is not necessarily the case. Due to coarticulation and free variation, different tokens of the same phoneme—even different parts of the same token—may be quite different in articulation. As a result, samples that are in fact from quite different articulatorily could be classified together as coming from the same oral configuration. Context-dependent distributions based on these populations would not accurately capture variability due to oral context, and would be unsuitable for normalization. Moreover, during measurement, the mapping of incoming acoustic samples to oral contexts would be unreliable.

A third problem with using phonemes to classify samples is that it assumes the vowel phoneme has already been identified in advance of nasality measurement. In speech recognition applications, the reverse might often be desired: knowing first whether a vowel is nasal may be used to help identify the vowel, since a different formant tracking technique might work better in nasal contexts (Fant, 1960; Glass and Zue, 1985).

An alternative characterization of oral context which avoids these drawbacks is position in formant space. The $F1$ and $F2$ values of a sample are directly related to the articulatory configuration of the oral cavity at a point in time, irrespective of phoneme category. Thus two samples which are close in formant space can perhaps more safely be considered to be articulatorily similar than two samples which merely have the same phoneme affiliation.

Unlike the phoneme inventory, formant space does not inherently divide samples into discrete groups. Therefore a different strategy must be adopted for acquiring means and standard deviations of the parameters in local oral contexts. For this purpose, we sample the formant space in an $n \times n$ grid pattern as shown in Figure 10. The grid vertices are evenly spaced along each formant's axis in the range from -3 to $+3$ standard deviations from the mean. Each grid point is a station to which a local mean and standard deviation will be assigned. However, rather than computing the mean and standard deviation of a grid point from a subset of the sample population, as was done with the phoneme contexts, they will be computed from the entire population. Nonetheless, the mean and standard deviation at a grid point will be local values by virtue of basing them most heavily on samples located closest to the grid point in formant space. This drop-off of influence with distance from the grid point is accomplished by assigning a weight to each sample based on its Euclidean distance in formant space from the grid point. Figure 11 illustrates the weighting of samples attenuating with distance from a grid point.

Various weighting functions may be used. A simple choice is the inverse of distance. Formally, for a given grid point g, the weight of each sample i with respect to g is the inverse of its distance from g (normalized so that all weights sum to one):

$$weight(g, i) = \frac{1/dist(g, i)}{\sum_j 1/dist(g, j)} \tag{2}$$

For each nasality parameter, a local mean and standard deviation are computed at each grid point based on the sample weightings with respect to that grid point. Specifically, the mean of parameter p at grid point g is the *weighted* mean of the

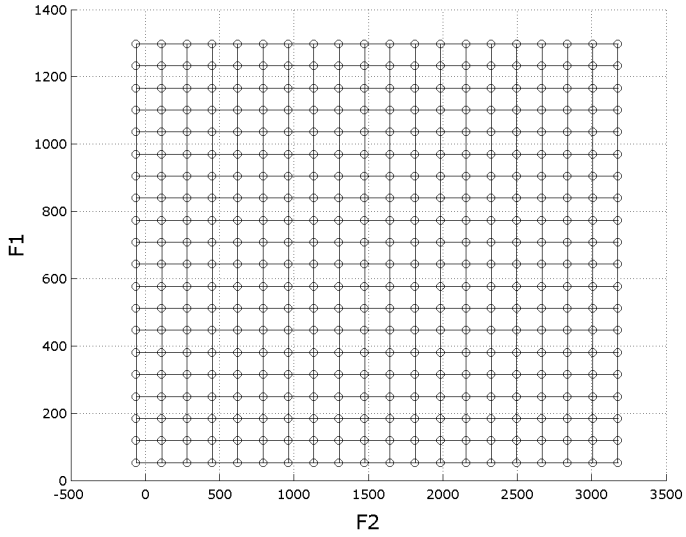

Figure 10: Sampling of the formant space in a grid pattern. Vertices of the grid (indicated by circles) are evenly spaced in each dimension over the range from -3 to $+3$ standard deviations from the mean.

parameter over all sample points.

$$\mu(p,g) = \sum_i weight(g,i)p(i) \tag{3}$$

After computing the mean of p at g, the variance of p at g can be derived as the *weighted* mean of squared deviations from the mean at g; the standard deviation at g can then be derived from the variance.

$$\sigma^2(p,g) = \sum_i weight(g,i)(p(i) - \mu(p,g))^2 \tag{4}$$

$$\sigma(p,g) = \sqrt{\sigma^2(p,g)} \tag{5}$$

Figure 12 is a plot showing the local means and standard deviations at the grid points for the parameter $COG(1000)$ for speaker B11. Three surfaces are

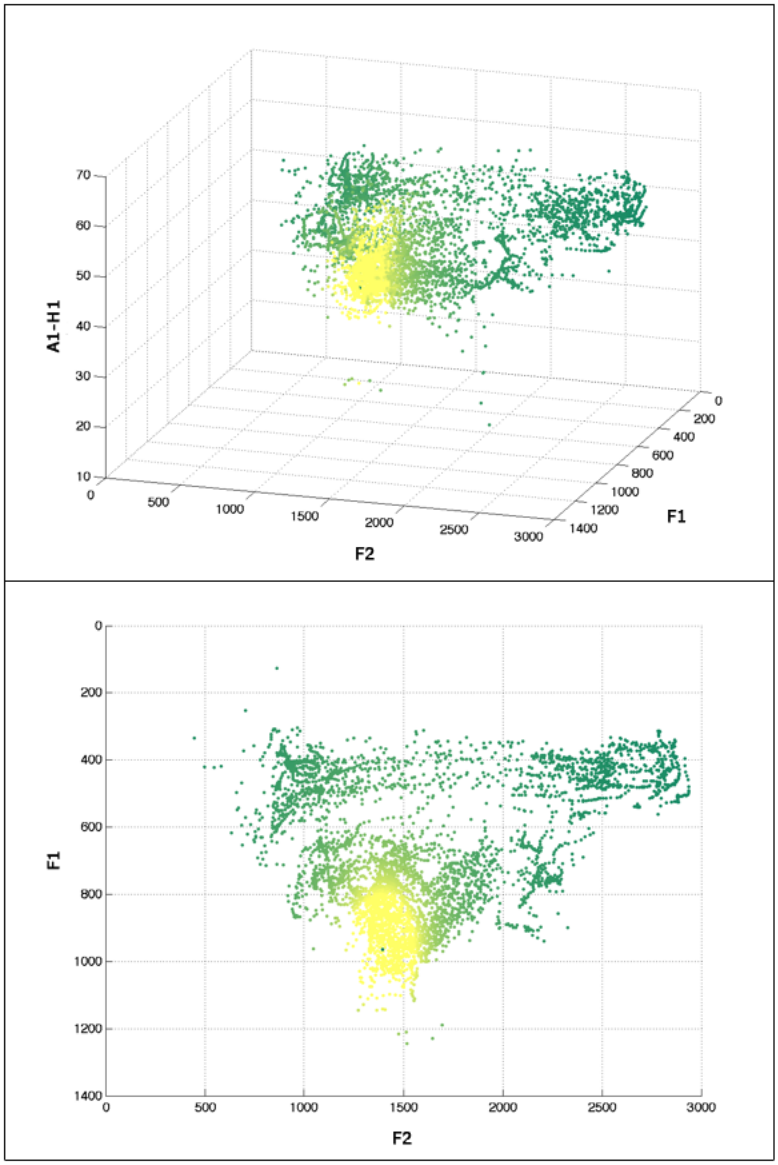

Figure 11: *Top:* Sample set of female Bengali speaker B11 plotted in formant space and in one of the nasality dimensions ($A1 - H1$). Coloring shows the weighting of samples based on distance in formant space from a particular grid point. The coordinates of the grid point are approximately 976 Hz (F1) and 1395 Hz (F2). Samples closer to the grid point in formant space will have more weight in determining the local mean and standard deviation of the nasality parameter at the grid point. *Bottom:* The same samples and weighting displayed in formant space only.

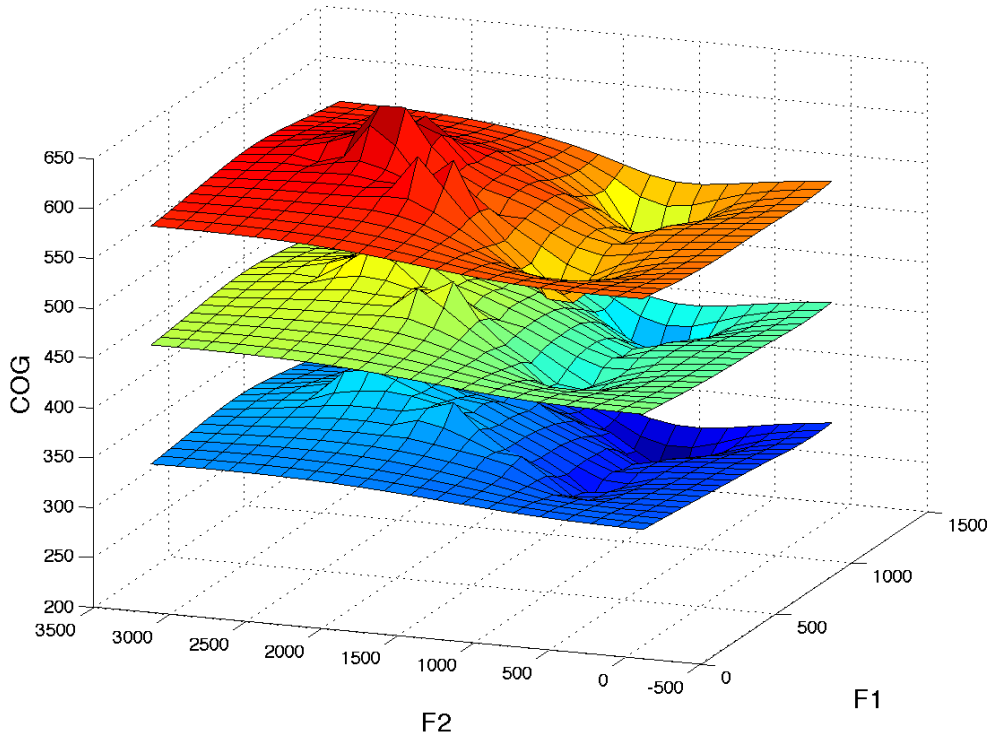

Figure 12: Means and standard deviations of a nasality parameter (COG within 1000 Hz) over formant space for a female Bengali speaker (B11). The central green surface represents bilinear interpolation of the local means of COG at the grid points in formant space. The top red surface and bottom blue surface are bilinear interpolations of the grid-point means plus and minus the corresponding grid-point standard deviations, respectively.

shown. The center surface is a bilinear interpolation of the grid-point means, and the top and bottom surfaces are interpolations of the grid-point means plus and minus the corresponding grid-point standard deviations, respectively. This plot is analogous to the plot in Figure 9 of means and standard deviations by vowel phoneme: each grid point serves a similar function to one of the phonemes. This plot clearly illustrates the deforming of the nasality parameter distribution over formant space.

As illustrated by the surfaces in Figure 12, local mean and standard deviation values for the nasality parameter may obtained at intermediate points in formant space by way of bilinear interpolation of the values at the grid points. Thus the context-dependent distribution of the parameter is defined continuously over formant space. Within the speaker model, $\mu_p(\vec{f})$ and $\sigma_p(\vec{f})$ represent the local mean and standard deviation, respectively, of a nasality parameter p at a point $\vec{f}$ in formant space. We may state the normalization function analogously to that used for phoneme-dependent distributions. For a sample with formant-space coordinates $\vec{f}$, its value x in nasality parameter p is adjusted by the z-score

$$N(p, x, \vec{f}) = \frac{x - \mu_p(\vec{f})}{\sigma_p(\vec{f})} \tag{6}$$

This is identical in form to equation 1, the only difference being that the oral context is now a vector in formant space rather than a vowel phoneme. Note that this normalization function, like the one in equation 1, can be applied to new data from the same speaker; the data need not be phonetically balanced, and—in contrast to normalization using phonemes—phoneme labels on the vowel intervals are not required.

A potential problem with using distributions dependent on position in formant space is that $F1$ may not always be a reliable indicator of oral configuration. It is generally agreed that $F1$ is affected not just by vowel articulation but also by nasal coupling: nasalization is said to cause $F1$ to shift upward (House and Stevens, 1956). Due to the effect of nasality itself on $F1$, the proximity of samples in formant space may actually not be such a reliable indicator of closeness in oral configuration. This then raises the same concerns as normalization based on phonemes: namely, that local distributions may not accurately capture

variability due to oral context, and that during measurement, the mapping of incoming acoustic samples to oral contexts may be unreliable. In Section 6.2, the two types of contextual distributions—one dependent on vowel phonemes and the other dependent on position in formant space—will be comparatively evaluated in terms of how well they enable the normalization function to reduce the variability in the parameters due to oral context.

6.3 Deriving the parameter integration function

After acquiring the context-dependent distributions of the nasality parameters for a speaker (based on either conception of oral context), the final step in the training algorithm is to define how to integrate the various nasality parameters into a single measure of nasality. A commonly used technique for dimensionality reduction is Principal Components Analysis (PCA). PCA produces a transformation that aligns the dimension of greatest variance with the first coordinate axis, the (orthogonal) dimension of second greatest variance with the second coordinate axis, etc. In the current approach, the parameter integration function is based on a PCA transformation of the nasality parameter space. Only the first principal component is used: after transformation, the value of an acoustic sample along the primary axis is considered its degree of nasality.

The PCA is computed from the entire sample set of the speaker's training data. Prior to PCA, the parameter values of the samples are first normalized using the normalization function (equations 1 or 6).

One cannot anticipate in advance of the PCA whether the first principal component will be configured to increase or decrease with nasality. For clarity it may be desirable to reverse the nasality dimension, so that it is oriented to

increase with nasality. This orientation correction may be accomplished automatically, using the following method. (1) Classify the speaker's training data samples into those taken from oral contexts (CV(C)) and those taken from nasal contexts, (C$\tilde{\text{V}}$N and C$\tilde{\text{V}}$(C)); (2) take the mean nasality value for each of the two populations; (3) if the oral mean is greater than the nasal mean, invert the transformation matrix.

As with other aspects of the speaker model, the parameter integration function (i.e., PCA transform) is speaker-specific.

7 Measuring algorithm

The speaker model generated by the training algorithm specifies how to measure nasality for a particular speaker. It defines the nasality parameters to use, their distributions in different oral contexts of the speaker, and their contributions to an integrated measure of nasality. This model can be used to measure nasality at points in time either in the training data or in a set of new vowel tokens recorded from the same speaker. Given a set of input acoustic data, the main steps for performing these measurements—the measuring algorithm—are as follows.

1. Sample and parameterize the vowels. As in the training algorithm, sample at glottal pulse extrema. For each sample, obtain values for each of the nasality parameters included in the model. If the parameter distributions in the model are defined to be dependent on position in formant space, then also obtain values for $F1$ and $F2$.

2. For each sample, normalize its nasality parameter values using the normalization function (equation 1 or 6, depending on the type of oral context

the parameter distributions are dependent on.)

3. For each sample, reduce the vector of parameter values to a single nasality measure by applying the parameter integration function (i.e. take the first principal component of the PCA transform). Due to the z-scores in the preceding step, the nasality measure will be centered around zero and given in units of standard deviation.

The acoustic data to which the measuring algorithm is applied need not be phonetically balanced; however, the vowels must still be marked off in some way and, if normalizing based on vowel classes, they must have phoneme labels as well.

Figures 13-15 demonstrate measurements of nasality over time using the measuring algorithm as just described. The speaker model used for these measurements included the nasality parameters $A1 - H1$ and $COG(1000)$, and parameter distributions based on formant space locations. Each figure plots the nasality measure (first principal component of the PCA) as a function of time for multiple tokens of a vowel spoken by one speaker in two or three contexts: oral, contextually nasal, and (Bengali only) contrastively nasal.

8 Evaluation

Ideally, to evaluate the success of the measuring system in capturing a dimension correlating with velar position, one would directly compare the acoustic-based measurements it produces with more direct measurements of velar position, such as those discussed in Section 1.3.1. This type of evaluation would require simultaneous recording of acoustic and articulatory data. However, synchronized

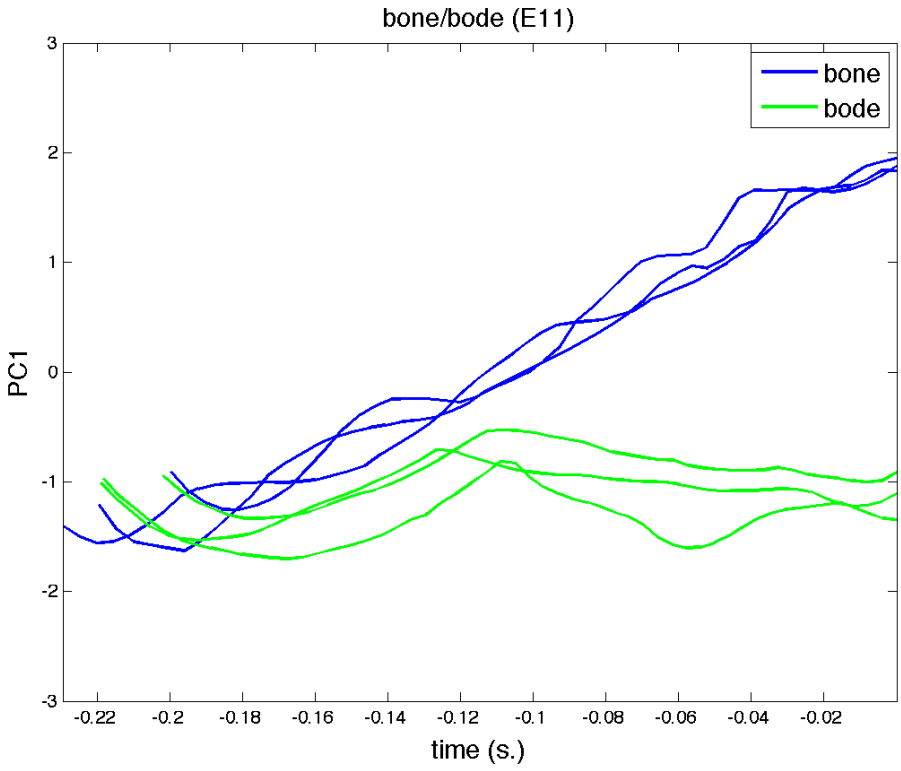

Figure 13: The nasality measure (first principal component of the PCA) as a function of time for multiple tokens of the English vowel /oʊ/ spoken by E11 in two contexts: nasalized in the word /boʊn/ (shown in blue) and oral in the word /boʊd/ (shown in green). Only the vowel intervals are shown. Vowel offsets are aligned at time 0.

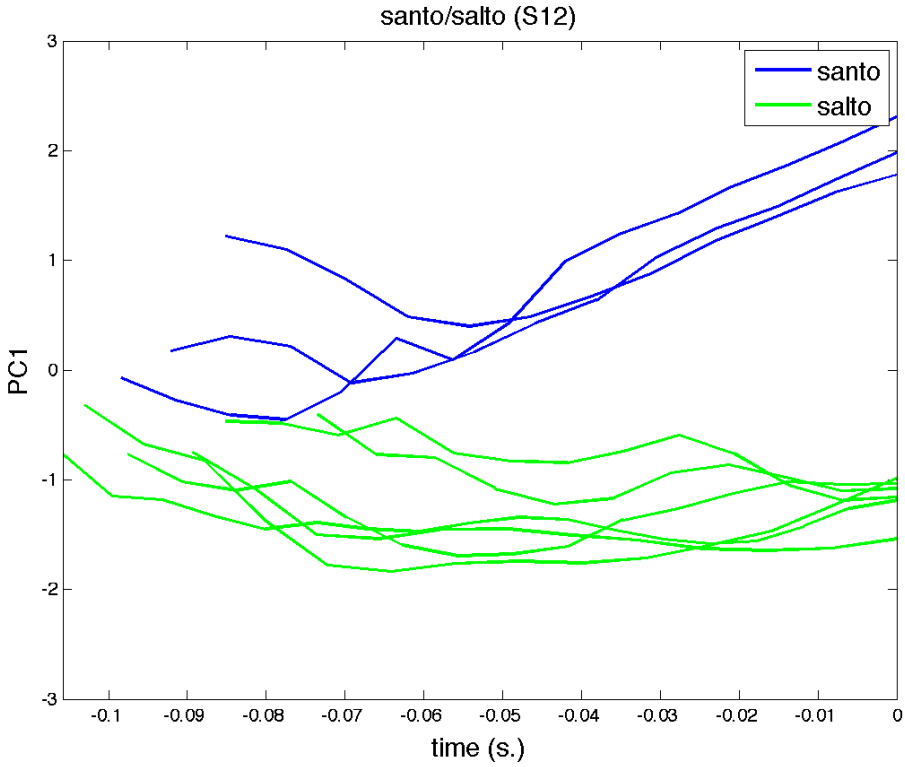

Figure 14: The nasality measure (first principal component of the PCA) as a function of time for multiple tokens of the Spanish vowel /a/ spoken by S12 in two contexts: nasalized in the word /santo/ (shown in blue) and oral in the word /salto/ (shown in green). Only the vowel intervals are shown. Vowel offsets are aligned at time 0.

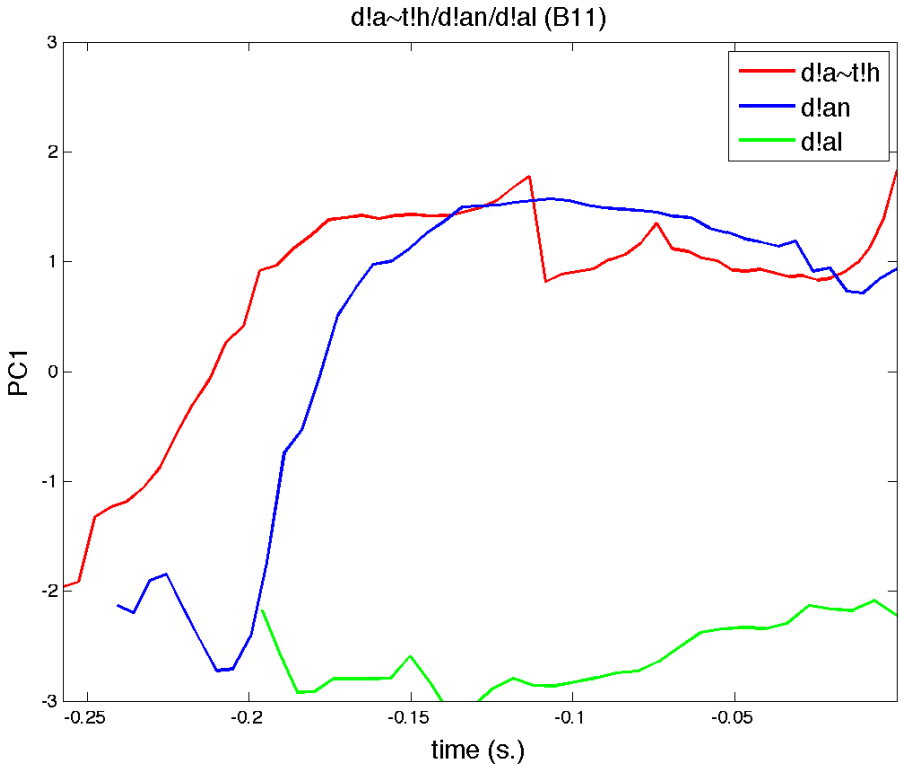

Figure 15: The nasality measure (first principal component of the PCA) as a function of time for tokens of the Bengali vowel /a/ spoken by B11 in three contexts: contrastively nasalized in the word /ɖãʈʰ/ (red line), contextually nasalized in the word /ɖan/ (blue line), and oral in the word /ɖal/ (green line). Only the vowel intervals are shown. Vowel offsets are aligned at time 0.

articulatory data was not available with the acoustic data used in this study.

Alternatively, one could look for confirmation of the acoustic measurements in some measure of perception of nasality. This would require design of a perceptual study in which the perception of nasality in a stimulus could be quantified as a function of time (see Section 1.3.2). Then, at each time point in the stimulus, the perceptual measure acquired from one or more subjects could be compared to an acoustic measure produced by the measuring system. A perceptual study of this nature is currently under development here at the University of Rochester.

Lacking reference to simultaneous articulatory or perceptual data, one may evaluate the measuring system in more indirect ways. In particular, one may look at statistical properties of the measurements to see how well they fulfill certain expectations. However, even with a scoring method of this type, any evaluation of a measuring system is somewhat meaningless unless we can compare it to another measuring system to or a gold standard. In the present case, the measuring system will not be compared to other systems, but rather, the system will be compared to itself using different options for components of the speaker model, or when possible it will be compared to worst-case performance, or best-case performance.

First, the various possible nasality parameters will be comparatively evaluated based on the criteria of discrimination and average acceleration. Next, parameter normalization will be evaluated by comparing variability due to oral context in three cases: normalization using phoneme-based distributions, normalization using formant-based distributions, and the worst case of no normalization. Finally, parameter integration will be evaluated by comparison to the best case of capturing all variance in one dimension.

Several of the evaluation criteria used below make reference to an *a priori* classification of vowels as oral or nasal. "Oral" vowels are defined as phonemically oral vowels in an oral context (CV(C)), and "nasal" vowels include both phonemically nasal vowels (C$\tilde{\text{V}}$(C)) and phonemically oral vowels in a nasal context (C$\tilde{\text{V}}$N).

8.1 Evaluation of the nasality parameters

This section will comparatively evaluate the six nasality parameters described in Section 6.1, on the basis of two performance metrics: discrimination and average acceleration. The outcome of this evaluation may help determine the relative suitability of the nasality parameters for inclusion in the speaker models.

Discrimination: Ability of the parameter to discriminate between (pre-classified) nasal and oral vowels; or in other words, how good the parameter is at discriminating between contexts that tend to differ in velar position. Discrimination of a parameter p for a speaker s is calculated as follows:

$$DISC(p,s) = \frac{|\mu_o(p,s) - \mu_n(p,s)|}{\sigma(p,s)} \tag{7}$$

where $\mu_o(p,s)$ is the mean value of parameter p over all samples taken from oral vowel tokens of speaker s, $\mu_n(p,s)$ is the mean value over all samples taken from nasal vowel tokens of speaker s, and $\sigma(p,s)$ is the standard deviation of p over all samples of speaker s (from both oral and nasal vowels). A higher discrimination value is better.

Average acceleration: Average absolute acceleration of the parameter. In general, the dynamics of a nasality parameter should be a believable reflection of velar dynamics. Given that the velum is a fairly slow moving articulator, the parameter should be penalized for abrupt changes over time or changing direction several times during the course of a vowel. Such behavior is probably indicative of noise in the parameter since it is unlikely to be due to the velum. In other words, an overall low amount of acceleration (positive or negative) is preferred. Average (absolute) acceleration of a parameter p for a speaker s is approximated as follows. For each vowel token v spoken by s, for each pair of successive samples $(i, i+1)$ in v, a velocity value is computed as

$$vel(v, i, i+1) = \frac{\Delta_p(v, i, i+1)}{\Delta_t(v, i, i+1)} \tag{8}$$

where $\Delta_p(v, i, j)$ is the change in the value of parameter p between samples i and j of p, and $\Delta_t(v, i, j)$ is the difference in time between the two samples. Next an absolute acceleration value is computed for each sample based on the two incident velocities:

$$acc(v, i) = \frac{|vel(v, i-1, i) - vel(v, i, i+1)|}{\Delta_t(v, i-1, i+1)} \tag{9}$$

Finally, the average acceleration of parameter p for speaker s is computed as the mean of the accelerations measured at all the samples of all the vowel tokens of s:

$$ACC(p, s) = \frac{\sum_{v, i \in v} acc(v, i)}{\sigma(p, s) N_s} \tag{10}$$

where N_s is the total number of samples of speaker s, and $\sigma(p, s)$ is the standard deviation of parameter p over all samples of speaker s. The metric is given in

Speaker	$A1-H1$	$COG(1000)$	$COG(1500)$	$B1$	$A1-P0$	$A1-P1$
B11	1.0354	0.4062	0.1032	0.6673	0.2478	0.5490
B12	0.6025	0.6176	0.2478	0.7175	0.6357	0.4026
B13	0.5084	0.2199	0.2998	0.4424	0.0993	0.6413
B14	0.0647	0.1158	0.2776	na	na	na
E11	1.0135	0.5449	0.4471	0.1183	0.7624	0.2631
E12	0.8570	0.5567	0.2215	0.4141	0.5453	0.4937
E13	0.5183	0.8580	0.7686	0.5030	0.7271	0.4251
E14	0.7624	0.6762	0.6426	0.0896	0.2525	0.3354
E15	0.3777	0.8170	0.8003	0.3166	0.8760	0.1264
E16	1.0134	1.0303	0.8473	0.6494	1.0900	0.4883
S11	0.6578	0.4196	0.1671	0.5620	0.4987	0.3893
S12	1.1948	0.6009	0.4854	0.8497	0.6176	0.2994
S13	0.6296	0.1457	0.1342	0.1818	0.1852	0.3642
S14	0.3473	0.1287	0.0759	0.4711	0.3452	0.0814
S15	0.8344	0.3781	0.2816	0.3414	0.4884	0.1549
S16	0.4913	0.0005	0.0513	na	na	na
S17	0.9859	0.1388	0.1246	1.0978	0.5503	0.5879
Average	0.6997	0.4503	0.3515	0.4948	0.5281	0.3735

Table 3: Discrimination results for each nasality parameter by speaker and averaged over speakers.

units of standard deviations per second per second.

Table 3 presents the discrimination results for each of the six nasality parameters based on the training data, by speaker and averaged over speakers. Looking at the averages over all speakers at the bottom of Table 3, $A1-H1$ had the best overall discrimination score, $A1-P0$ was second, $B1$ and $COG(1000)$ were both middling, while $A1-P1$ and $COG(1500)$ had relatively poor discrimination. Notably, $COG(1000)$ had considerably better discrimination than $COG(1500)$.

Table 4 shows the results in average acceleration for each parameter. Again looking at cross-speaker averages, the worst overall performers were $B1$ and $A1-P1$. In contrast to its good discrimination results, the parameter $A1-H1$

Speaker	$A1-H1$	$COG(1000)$	$COG(1500)$	$B1$	$A1-P0$	$A1-P1$
B11	1408.3	849.1	712.6	4381.7	1852.1	1860.2
B12	2984.2	1355.9	1233.6	3060.0	2579.3	2596.5
B13	2262.5	1055.1	1187.4	1850.4	1229.3	2233.6
B14	2646.3	1801.9	1761.1	na	na	na
E11	1300.4	582.2	574.4	5880.3	1515.6	1313.7
E12	1587.7	1046.3	942.6	4917.4	2281.1	1768.1
E13	1630.3	930.8	897.6	3130.4	1828.0	1177.4
E14	1125.7	786.3	784.8	2341.3	1049.1	1076.5
E15	2671.4	993.5	990.4	3923.2	2281.9	2224.2
E16	2021.4	1311.0	1314.7	4718.7	2311.7	1741.1
S11	4014.4	998.4	876.1	3138.4	3271.9	2076.6
S12	1363.0	794.7	780.2	1635.1	1472.8	1717.7
S13	2673.5	1361.3	1713.6	3607.5	2404.7	2946.3
S14	2442.5	1267.3	1434.6	3247.3	2737.8	3006.4
S15	2278.6	1327.9	1410.7	3467.2	1672.1	3136.3
S16	1694.3	1092.0	1215.9	na	na	na
S17	2304.9	1189.3	1241.8	1836.6	1494.2	1848.7
Average	2141.7	1102.5	1121.9	3409.0	1998.8	2048.2

Table 4: Average acceleration results (in standard deviations per second per second) for each nasality parameter, by speaker and averaged over speakers.

also had high average acceleration, indicating that this is a relatively jumpy parameter even though it is sensitive to nasality. COG—in both the 1000 and 1500 Hz ranges—had the least acceleration overall (that of $COG(1000)$ was slightly better).

To better illustrate differences in acceleration between parameters, Figure 16 shows measurements of $A1 - H1$ and $COG(1000)$ (after normalization) over the time course of one token of the vowel /a/ in *gone*, spoken by speaker E15. Note in Table 4 that for this speaker, $A1 - H1$ has an above-average score for average acceleration, and $COG(1000)$ has a below-average score. The plot in Figure 16 exemplifies this difference in acceleration: although the two parameters more or less track each other, $A1 - H1$ displays a more erratic, noisy sort of behavior. Of course, this figure only illustrates one vowel token and does not necessarily show how the parameters behave generally for this speaker.

The results of the discrimination and average acceleration tests may be studied in combination, to weigh the general suitability of each parameter. Looking at cross-speaker averages, the two metrics do not always agree in their selection. Although $B1$ had a moderate discrimination score, its exceptionally high acceleration score indicates that it is probably too noisy a parameter to be useful. (The noise may perhaps be due to the bandwidth-measuring function in Praat, rather than an inherent instability in bandwidth. However, this is difficult to assess since formant bandwidth is not a rigorously defined concept independently of the algorithm used to measure it.) Similarly, $A1 - H1$ had the highest discrimination but scored poorly overall in average acceleration. $COG(1000)$ had worse-than-average discrimination but better-than-average acceleration. On certain comparisons the two metrics agree. For example, $A1 - P1$ scored poorly in both

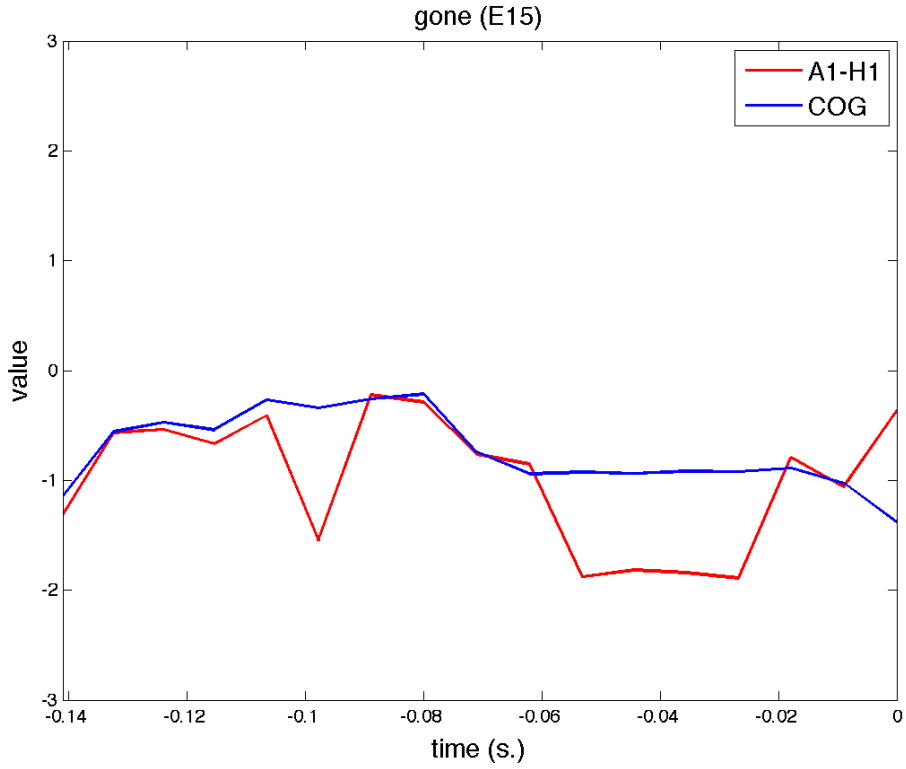

Figure 16: Measurements of $A1-H1$ (red) and $COG(1000)$ (blue) (after normalization) over the time course of one token of the vowel /a/ in *gone*, spoken by speaker E15. The comparison exemplifies the difference in acceleration between the two parameters for this speaker.

acceleration and discrimination, and $COG(1000)$ was superior to $COG(1500)$ in both metrics.

Perhaps more interesting than these general comparisons between cross-speaker averages, however, is the variability in the results between different speakers. For example, contrary to the general trends, for speaker B13, $A1-P1$ has the highest discrimination score rather than one of the worst, $COG(1500)$ has better discrimination than $COG(1000)$, and $B1$ has a better-than-average acceleration score. What this may indicate is that parameters that perform best on average may not necessarily be the most suitable ones for individual speakers.

We can use this observation to our advantage. To improve the measuring performance of the system for each speaker, each speaker model could be tailored include the best-scoring parameters for that speaker. Alternatively, a weighting could be assigned to the parameters based on the scores, and these weightings could be used to impart more or less influence to individual parameters in the parameter integration function.

In a future stage of this work, the comparative evaluation of the nasality parameters may be automated and incorporated into the training algorithm. For each speaker's training, a large set of nasality parameters may be automatically ranked using these criteria to determine which are the most reliable indicators of nasality for that speaker. The results of the ranking would be used to either select or weight the nasality parameters to be included in the speaker model.

In principle, speaker-specific parameter selection constitutes another way of normalizing nasality measurement over speakers: by emphasizing in the speaker model the parameters that seem most correlated with velar position for that speaker, we effectively encapsulate these speaker-specific manifestations of nasal-

ity, leaving an externally normal measure.

8.2 Evaluation of parameter normalization

The purpose of parameter normalization is to remove variability in the nasality parameters due to speaker and oral context. The context-dependent parameter distributions in the speaker model serve to capture that variability so that it can be removed. The two types of distributions described in Section 6.2—one dependent on vowel phonemes and the other dependent on position in formant space—may thus be comparatively evaluated in terms of how well they enable the normalization function to reduce the variability in the parameters due to oral context.

One way to estimate variability due to oral context is to attempt to minimize the variability due to nasality. We may minimize the effect of nasality by restricting attention to samples taken from vowels of the same nasality class (oral or nasal). The following metric estimates variability due to oral context as variability within nasality class relative to total variability.

Deviation within nasality class: A measure of deviation between samples taken from the same nasality class—oral or nasal. The *oral* deviation of a parameter p for a speaker s, $DEV_o(p, s)$ is defined as the ratio of the standard deviation of samples taken from oral vowels to the standard deviation of the whole sample population. Similarly, the *nasal* deviation of p for s, $DEV_n(p, s)$, is the ratio of the standard deviation of samples taken from nasal vowels to the

standard deviation of the whole population.

$$DEV_o(p,s) = \frac{\sigma_o(p,s)}{\sigma(p,s)} \qquad DEV_n(p,s) = \frac{\sigma_n(p,s)}{\sigma(p,s)} \tag{11}$$

The total deviation measure is the average between oral and nasal deviations.

$$DEV(p,s) = \frac{DEV_o(p,s) + DEV_n(p,s)}{2} \tag{12}$$

Smaller deviation values imply the presence of less variability due to oral context. Deviation scores were computed for the nasality parameters $A1 - H1$ and $COG(1000)$ based on the training data both before and after applying normalization. Both types of normalization were evaluated. For the case of normalization based on position in formant space, the context-dependent parameter distributions were constructed on the basis of $20 \times 20 = 400$ grid points in formant space. (This density of grid points seemed to strike the right balance between processing/memory load and the ability to capture local distributions.)

Tables 5 and 6 report deviation within nasality class for the parameters $A1 - H1$ and $COG(1000)$, respectively, for three cases: the case in which the parameter is normalized based on phonemes (the "Phoneme" column); the case in which it is normalized based on formant space (the "Formant" column); and to provide a worst case, the case in which the parameter is not normalized (the "None" column). The scores in this table may be interpreted as percentages. For example, the value 0.89 in the "None" column for speaker B11 in Table 5 means that without normalization, the standard deviation of $A1 - H1$ among samples within the same vowel class is 89% of the standard deviation over the entire sample population of the speaker. The goal of normalization is to bring

Speaker	None	Phoneme	Formant
B11	0.89	0.84	0.88
B12	0.94	0.95	0.96
B13	0.96	0.96	0.96
B14	0.99	1.01	1.00
E11	0.86	0.79	0.84
E12	0.90	0.87	0.88
E13	0.94	0.93	0.91
E14	0.92	0.91	0.92
E15	0.97	0.96	0.97
E16	0.86	0.84	0.84
S11	0.93	0.91	0.91
S12	0.79	0.77	0.79
S13	0.94	0.93	0.94
S14	0.99	0.99	0.99
S15	0.91	0.88	0.91
S16	0.98	0.95	0.96
S17	0.87	0.85	0.88
Average	0.92	0.90	0.91

Table 5: Results of deviation within nasality class for $A1 - H1$ with no normalization, normalized over phonemes, and normalized over formant space, by speaker and with cross-speaker averages.

these percentages down.

It can be seen in Tables 5 and 6 that for many of the speakers, the normalization functions do bring the deviation scores down relative to the case with no normalization. However, the goal is achieved only in small measure. As indicated by the values at the bottom of each table, on average the reduction in deviation is only one or two percentage points.

The apparent failure of the measuring system to normalize over oral context may be considered further in Figure 17. This figure shows an average, time-normalized nasality contour for each vowel phoneme of speaker E11, generated by averaging over only the *oral* tokens of each vowel. (These measurements

Speaker	None	Phoneme	Formant
B11	1.02	0.95	0.98
B12	0.98	0.96	0.96
B13	0.99	0.96	0.98
B14	1.01	1.01	1.01
E11	0.96	0.87	0.91
E12	0.96	0.93	0.92
E13	0.90	0.86	0.88
E14	0.93	0.93	0.93
E15	0.90	0.88	0.91
E16	0.85	0.84	0.84
S11	0.94	0.92	0.91
S12	0.93	0.91	0.90
S13	0.99	0.99	0.98
S14	0.99	1.00	0.99
S15	0.99	0.92	0.96
S16	1.01	1.01	1.00
S17	0.99	0.99	0.98
Average	0.96	0.94	0.94

Table 6: Results of deviation within nasality class for $COG(1000)$ with no normalization, normalized over phonemes, and normalized over formant space, by speaker and with cross-speaker averages.

were produced using normalization based on formant space.) To the extent that normalization over oral context is effective, we expect to see convergence between tokens of different vowel phonemes that are in the same nasality class. The scattering of vowels over the range from −2 to 2 standard deviations would seem to show that this convergence is not taking place.

However, on closer inspection of the graph in Figure 17, it can be seen that there is a concentration of vowels in a band between −2 and 0, and three outlying vowels between 1 and 2. These outlying vowels are the high vowels /i/, /u/, and the diphthong /ju/ (in the figure key these are represented by the ARPABET symbols /iy/, /uw/ and /yuw/, respectively). Thus, if we exclude high vowels from consideration, it may be found that normalization works better than indicated by the deviation results.

Figure 18 further confirms the exceptional status of high vowels. This figure presents the same comparison between vowels, but this time each curve is an average over the *nasalized* tokens of each vowel for speaker E11. Here the most outlying vowels are /i/, /u/, the diphthong /ju/, and (only at the tail end of the vowel interval), the dipthongs /eɪ/, /ɔɪ/, /aɪ/, and perhaps /aʊ/ and /oʊ/. (In the key, these are symbolized by /iy/, /uw/, /yuw/, /ey/, /oy/, /ay/, /aw/ and /ow/). Again, these are all high vowels or diphthongs that end in high vowels.

In sum, the failure of the normalization function to eliminate the effect of oral context on nasality measurement seems to be mostly a failure to eliminate the effect of vowel height. This is not an effect limited to speaker E11 but was seen across the speakers of this study. This problem will be discussed in more detail in Section 9.

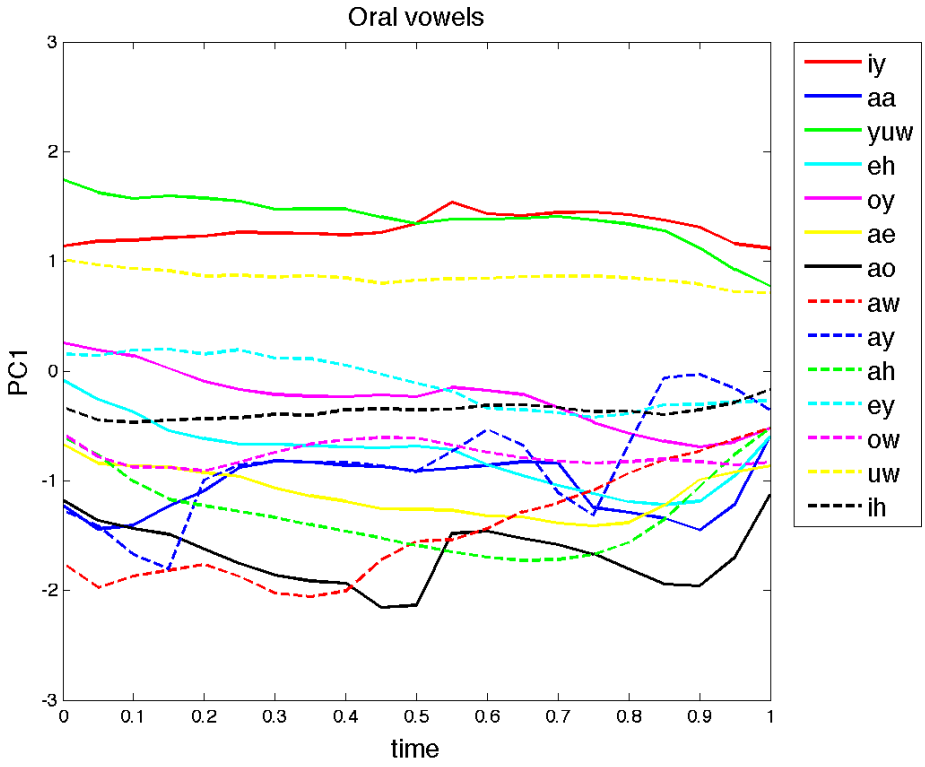

Figure 17: Nasality contours of different oral vowels of speaker E11, averaged over tokens and time-normalized.

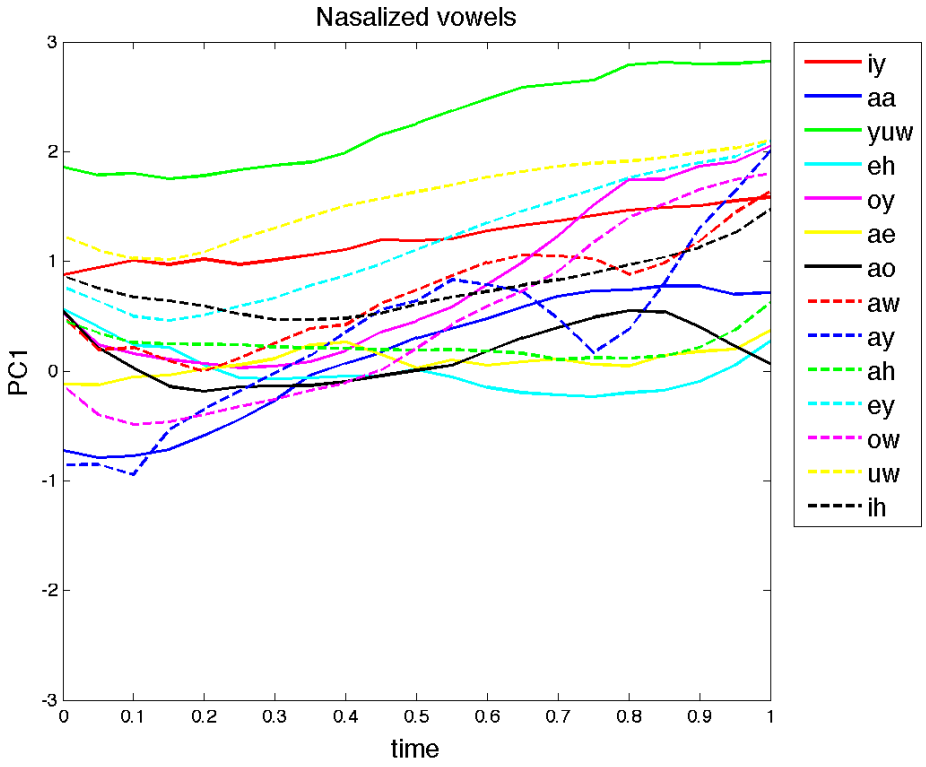

Figure 18: Nasality contours of different nasalized vowels of speaker E11, averaged over tokens and time-normalized.

8.3 Evaluation of parameter integration

The final aspect of the measuring system that may be evaluated is parameter integration, implemented here as PCA. The aim of parameter integration is to reduce measurements in multiple parameters to one parameter—representing a measure of degree of nasalization—without excessive loss of information. Since the first principal component of the PCA is used as the nasality measure, the best-case result would simply be to capture 100% of the variance in the first component. Thus the success of parameter integration may be rated in terms of how close the first component gets to that target. The proportion of variance accounted for by the first principal component is also indicative of how closely the nasality parameters in the speaker model are correlated with each other.

Table 7 lists by speaker the percentage variance captured by the first principal component using a speaker model with two nasality parameters $A1-H1$ and $COG(1000)$. The first principal component captures on average 85.9% of the variance. The implication is that these two parameters are fairly well correlated.

9 Discussion and future work

In the past, researchers attempting to measure nasality acoustically have generally looked for acoustic correlates that influence *perception* of nasality, as when manipulated in synthetic speech. It has been a theme of this paper that measurement of vowel nasalization is not only a study of a perceptual factor, but it is also a form of articulatory recovery. There are several reasons for viewing nasality measurement in vowels as a method of recovering velar position. First, nasality in vowels is caused by velar lowering. Causes can often be inferred from

Speaker	PC1
B11	88.7
B12	80.0
B13	84.9
B14	82.0
E11	89.5
E12	90.5
E13	89.1
E14	87.8
E15	81.7
E16	92.6
S11	75.6
S12	89.0
S13	82.6
S14	86.0
S15	88.0
S16	89.7
S17	82.9
Average	85.9

Table 7: Percentage variance captured by the first principal component for each speaker using a model with two nasality parameters $A1 - H1$ and $COG(1000)$.

their effects. Therefore, it is reasonable to suppose that degree of velar lowering can be inferred from degree of acoustic nasality. Secondly, and perhaps more importantly, it is difficult or impossible to ignore causal articulatory factors when attempting to measure nasality. This is because nasality varies not only as a function of velar position but also due to differences in speaker and oral cavity shape. Attempts to isolate a single acoustic dimension of nasality are thwarted unless one can normalize acoustic measurements across these extra articulatory factors. In other words, even if one has no interest in the the articulatory causes leading to nasality, the effects corresponding to those causes must still be separated from each other if one is to access a coherent nasality dimension. That coherent dimension is thus intrinsically tied to the articulatory dimension that formed it.

The extent to which it is actually possible to measure nasality and recover velar position, given the conflation of effects in the acoustics, is another question. It seems likely that some acoustic dimension corresponding to nasalization is accessible to human listeners. Nasalization is used contrastively in some speech communities; therefore, members of those communities must be able to detect differences in the level of nasality in vowels. However, this only means that that listeners have to decide between two poles: nasal and non-nasal; it does not necessarily imply the existence of a well-defined acoustic continuum on which we can base measurements and from which we can infer velar lowering. Moreover, the ability to distinguish between nasal and non-nasal counterparts of a vowel—for example, between French *beau* /bo/ and *bon* /bõ/—may not even require or imply the independent perception of the vowel quality and the nasal feature. Rather, the nasalized vowel as a whole may be perceived as a distinct vowel

quality—a gestalt. This would represent a disconnect between speakers' perceptions of nasalized vowels and their knowledge of how to produce them—out of distinct velic and oral gestures. However, the idea that speakers use different knowledge for processing and producing speech is not that unreasonable.

In spite of these questions about the existence of a measurable "nasality dimension" in the acoustic parameter space of vowels, the presence of spectral modifications that vary gradually with nasalization invite one to seek such a dimension. But in doing so, one must still isolate the modifications due to velar lowering from those due to other articulatory factors.

In the present approach, an attempt is made to solve this inverse problem using a normalization technique. It may be instructive to compare this method to more traditional approaches to articulatory recovery. In the past, attempts at articulatory recovery have generally used the analysis-by-synthesis model (Stevens, 1960). Given acoustic input, one tries to determine parameter values in a vocal tract model that would allow one to synthesize an acoustic signal with spectral properties matching those of the original signal (e.g., McGowan, 1994). In contrast, the method presented here analyzes articulation from acoustics without resynthesizing acoustics from articulation. Rather, articulation is inferred "directly" from acoustics, based on the statistical properties of the acoustic parameters. (The phrase "direct articulatory inference" has been offered as a possible description of this procedure by Richard S. McGowan, personal communication. Note that this use of the word "direct" simply means without resynthesis; it does not carry the same import as in "direct perception" (Fowler, 1986).)

A disadvantage of the normalization approach is that it requires a set of

phonetically balanced training data for each speaker so as to elicit accurate statistics about the parameters. In the analysis-by-synthesis approach, exposure to balanced training data is not required, but one does require a model of the speaker's vocal tract—which seems a significantly taller order.

Without time-aligned articulatory data with which to compare the acoustic nasality measurements, an assessment of the accuracy of this method of articulatory recovery remains out of reach. However, a metric was proposed (Section 8.2) to evaluate the effectiveness of the normalization function in reducing variability due to oral context. This is relevant to determining the success of the measuring system in isolating the effects of velar lowering from the effects of vowel articulation. The metric indicated that the normalization function did not do much to reduce variability due to oral context. The primary problem seemed to be that vowel height had a continued influence on the nasality measure.

It is not exactly clear why the normalization function failed to normalize the parameters over differences in vowel height. However, it is likely to have to do with the similarities in the acoustic effects of velar lowering and vowel height. Both of these articulatory adjustments affect the same frequency region—namely, the region around $F1$. Both nasal coupling and lowering of the tongue cause raising of F1. Nasal coupling also introduces nasal formants in the same region where $F1$ might occur in corresponding oral vowels, creating further similarity. Experiments have shown that the acoustic similarity of the two articulatory factors also causes them to be similar perceptually, which can lead to trading relations (Krakow *et al.*, 1988). Nasalization effectively induces a perceived vowel space contraction: low vowels are perceived as higher and high vowels are perceived as lower in the presence of nasalization (Wright, 1975,

1986).

The similarity between the acoustic effects of tongue height and velar lowering may be an instance of the one-to-many problem encountered in articulatory recovery, in which one acoustic result can be mapped to several articulatory configurations. In this case, certain spectral features in the low frequency range around $F1$ can indicate either tongue-height or velum-height adjustments.

The dependence of the nasality parameters on vowel height may be abated by selecting different nasality parameters. It might be better, for instance, to use a larger set of parameters than was used in the evaluations given above. One might even use a set of parameters that characterize the entire spectrum (e.g., Mel Frequency Ceptstral Coefficients), rather than target just those particular features theoretically associated with nasal coupling. Given this large set of parameters, then, the training algorithm could include in the model the subset which have the best scores in parameter evaluation (Section 8.1). The selection of parameters might turn out differently for different speakers. Such an approach may be tried in the future.

Another tactic might be to give more careful examination to the accuracy of formant tracking. Correctly measuring the formant frequencies for each sample is critical to normalization based on formant space; however, in the presence of nasalization, formant trackers are often confused by the presence of nasal formants. To see whether poor formant tracking is contributing to poor normalization, it may be worthwhile to manually edit the automatically tracked formants; any improvement in normalization that resulted would indicate the extent to which normalization problems are due to formant tracking.

Another important aspect of the current approach that could be improved is

the phonetic balancing of the word lists. Some vowel phonemes are underrepresented; this is especially true in the Spanish list, which has only one word pair containing /i/. Furthermore, the Bengali list is not quite balanced between oral and nasalized vowels. While each vowel in the English and Spanish word lists appears in an even number of oral and nasal contexts, the Bengali list has more nasalized vowels than oral vowels due to the use of triples which each contain two nasalized vowels (one contextual and one contrastive). In all, a more precise approach to phonetic balancing may be required.

Future work should also include a more direct evaluation method. As noted earlier, an acoustic measure of nasalization is truly untested until it is compared with actual articulatory measurements, or alternatively, with measurements of perception of nasality. Both types of evaluations are currently being planned.

10 Summary

This thesis approached the problem of acoustic nasality measurement in vowels as a case of articulatory recovery, or the "inverse problem": inference of the articulatory state of the vocal tract from the acoustic signal. Specifically, measuring nasality in vowels was viewed as intrinsically being an attempt to isolate the acoustic effects of one articulatory variable—velar position—from the simultaneous and often conflated effects of other articulatory variables, namely speaker anatomy and the configuration of the oral cavity determined by the vowel articulation. An automated measuring system was presented which is designed to measure nasality independently of speaker and oral context using a normalization procedure.

This measuring system consists of a training algorithm, which generates a

set of speaker-specific information called a speaker model, and a measuring algorithm, which uses the speaker model to measure nasality in vowels of that speaker. The speaker model specifies the set of nasality parameters used to measure nasality for the speaker, the distributions of those parameters dependent on oral context, and a function for integrating the parameters into a single measure. The training algorithm first samples and parameterizes the vowels in the training data. It then acquires the context-dependent distributions of the nasality parameters from this sample set, based on either of two conceptions of oral context: vowel phonemes or position in formant space. Finally, the training algorithm determines how to integrate the multiple nasality parameters into a single measure of nasality for the speaker, by applying Principal Components Analysis to the sample set. The parameter integration function utilizes the first principal component of the PCA transform.

Lacking an independent measure of velar position or perception of nasality, the measuring system was evaluated using statistical metrics to compare between different options in the speaker model and to compare performance against best-case and worst-case performance. Different nasality parameters were comparatively evaluated using the criteria of discrimination and average acceleration. It was noted that while there were overall trends favoring some parameters over others, different sets of parameters may be suitable for different speakers. Speaker-specific parameter selection based on these criteria was proposed as an additional step of the training algorithm. The normalization procedure was evaluated in terms of its success in reducing variability in the parameters due to oral context. Both types of normalization—based on phoneme-dependent and formant-dependent distributions—were assessed in comparison

to the worst case of no normalization. It was found that neither method was very successful in reducing variability of the nasality parameters due to differences in oral context, and this failure was mainly attributed to the inability to separate the effects of nasal coupling from the similar effects of vowel height. Future work will attempt to use a larger set of nasality parameters and perhaps more careful formant tracking to overcome this problem. More precise phonetic balancing of the word lists may also be required. Finally, this nasality measure must be confirmed by comparison to direct articulatory measurements of velar position or with measurements of perception of nasality.

References

Amelot, A. (**2004**). "An aerodynamic, fiberscopic, acoustic and perceptive study of nasal vowels of french", Ph.D. thesis.

Andruski, J. E., Blumstein, S. E., and Burton, M. (**1994**). "The effects of subphonemic differences on lexical access", Cognition **52**, 163–187.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (**1978**). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", Journal of the Acoustical Society of America **63**, 1535–1555.

Baken, R. J. (**1987**). *Clinical Measurement of Speech and Voice* (College Hill, Boston).

Berger, M. A., Clayards, M., Bardhan, N., and McDonough, J. (**2007**). "Temporal patterns of vowel nasalization in three languages", ICPhS Saarbrucken, submitted.

Boersma, P. and Weenink, D. (**2007**). *Praat: doing phonetics by computer* (Version 4.5.16) [Computer program]. Retrieved February 22, 2007, from http://www.praat.org/.

Brehm, F. E. (**1922**). "Speech correction", American Annals of the Deaf **67**, 361–370.

Chafcouloff, M. and Marchal, A. (**1999**). "Velopharyngeal coarticulation", in *Coarticulation: Theory, Data and Techniques*, edited by W. J. Hardcastle and N. Hewlett, 69–79 (Cambridge University Press, Cambridge).

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

Chen, M. Y. (**1997**). "Acoustic correlates of english and french nasalized vowels", Journal of the Acoustical Society of America **102**, 2360–2370.

Clumeck, H. (**1976**). "Patterns of soft palate nmovements in six languages", Journal of Phonetics **4**, 337–351.

Cohn, A. (**1990**). "Phonetic and phonological rules of of nasalization", Ph.D. thesis, UCLA.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, Gravenhage).

Fletcher, S. G. and Frost, S. D. (**1974**). "Quantitative and graphic analysis of prosthetic treatment for "nasalance" in speech", Journal of Prosthetic Dentistry **32**, 284–291.

Fowler, C. A. (**1986**). "An event approach to the study of speech perception from a direct-realist perspective", Journal of Phonetics **14**, 3–28.

Glass, J. R. (**1984**). "Nasal consonants and nasalized vowels: An acoustic study and recognition experiment", Master's thesis, Massachusetts Institute of Technology.

Glass, J. R. and Zue, V. W. (**1985**). "Detection of nasalized vowels in American English", in *Proceedings of ICASSP*, volume 4, 1569–1572.

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K., and Wang, T. (**2005**). "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop", in *Proceedings of ICASSP*, volume 5, 213–216.

Hattori, S., Yamamoto, K., and Fujimura, O. (**1958**). "Nasalization of vowels in relation to nasals", Journal of the Acoustical Society of America **30**, 267–274.

Hawkins, S. and Stevens, K. N. (**1985**). "Acoustic and perceptual correlates of the nonnasal-nasal distinction for vowels", Journal of the Acoustical Society of America **77**, 1560–1575.

Horiguchi, S. and Bell-Berti, F. (**1987**). "The velotrace: a device for monitoring velar position", Cleft Palate Journal **24**, 104–111.

Horii, Y. (**1980**). "An accelerometric approach to nasality measurement: a preliminary report", Cleft Palate Journal **17**, 254–261.

House, S. A. and Stevens, K. N. (**1956**). "Analog studies of the nasalization of vowels", Journal of Speech and Hearing Disorders **21**, 218–232.

Huffman, M. K. (**1990**). "The role of f1 amplitude in producing nasal percepts", Journal of the Acoustical Society of America **88**, S54.

Karnell, M. P., Linville, R. N., and Edwards, B. A. (**1988**). "Variations in velar position over time: A nasal videoendoscopic study", Journal of Speech and Hearing Research **31**, 417–424.

Klopfenstein, M. (**2006**). "Phonetic implementation of phonological categories: the case of contextual and contrastive vowel nasalization in ottawa", Master's thesis, Wayne State University.

Krakow, R. A., Beddor, P. S., Goldstein, L. M., and Fowler, C. A. (**1988**). "Coarticulatory influences on the perceived height of nasal vowels", Journal of the Acoustical Society of America **83**, 1146–1158.

Lahiri, A. and Marlsen-Wilson, W. (**1992**). "Lexical processing and phonological representation", in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, edited by G. J. Docherty and D. R. Ladd, 229–260 (Cambridge University Press, Cambridge).

Maeda, S. (**1982**). "The role of the sinus cavities in the production of nasal vowels", in *Proceedings of IEEE International Conference*, volume 2, 911–914.

McGowan, R. S. (**1994**). "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary results", Speech Communication **14**, 19–49.

McMurray, B., Clayards, M., Aslin, R. N., and Tanenhaus, M. K. (**2004**). "Gradient sensitivity to acoustic detail and temporal integration of phonetic cues", poster presented at the 147^{th} Meeting of the Acoustical Society of America, May 2004.

Moll, K. L. (**1960**). "Cinefluorographic techniques in speech research", Journal of Speech and Hearing Research **3**, 227–241.

Ohala, J. J. (**1971**). "Monitoring soft-palate movements in speech", Journal of the Acoustical Society of America **50**, 140.

Ohala, J. J. and Ohala, M. (**1995**). "Speech perception and lexical representation: The role of vowel nasalization in Hindi and English", in *Papers in Laboratory Phonology IV: Phonology and Phonetic Evidence*, edited by B. Connell and A. Arvaniti, 41–60 (Cambridge University Press, Cambridge).

Pruthi, T. (**2007**). "Analysis, vocal-tract modeling and automatic detection of vowel nasalization", Ph.D. thesis, University of Maryland.

Quigley, L. F., Shiere, F. R., Webster, R. C., and Cobb, C. M. (**1964**). "Measuring palatopharyngeal competence with the nasal anemometer", Cleft Palate Journal **1**, 304–313.

Rochet, A. P. and Rochet, B. L. (**1991**). "The effect of vowel height on patterns of assimilation nasality in French and English", in *Proceedings* 12^{th} *ICPhS*, volume 3, 54–57.

Shelton, R., Arndt, A. K. W., and Elbert, M. (**1967**). "The relationship between nasality score values and oral and nasal sound pressure level", Journal of Speech and Hearing Research **10**, 542–548.

Sloan, G. M. (**2000**). "Posterior pharyngeal flap and sphincter pharyngoplasty: The state of the art", Cleft Palate-Craniofacial Journal **37**, 112–122.

Solé, M. (**1992**). "Phonetic and phonological processes: the case of nasalization", Language and Speech **35**, 29–43.

Solé, M. (**1995**). "Spatio-temporal patterns of velopharyngeal action in phonetic and phonological nasalization", Language and Speech **38**, 1–23.

Stevens, K. N. (**1960**). "Toward a model for speech recognition", Journal of the Acoustical Society of America **32**, 47–55.

Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, Mass).

Stevens, K. N., Kalikow, D. N., and Willemain, T. R. (**1975**). "A miniature accelerometer for detecting glottal waveforms and nasalization", Journal of Speech and Hearing Research **18**, 594–599.

Stevens, K. N., Nickerson, R. S., Boothroyd, A., and Rollins, A. M. (**1976**). "Assessment of nasalization in the speech of deaf children", Journal of Speech and Hearing Research **19**, 393–416.

Ushijima, T. and Hirose, H. (**1974**). "Electromyographic study of the velum during speech", Journal of Phonetics **2**, 315–326.

Weiss, A. (**1954**). "Oral and nasal sound pressure levels as related to judged severity of nasality", Ph.D. thesis, Purdue University.

Wright, J. T. (**1975**). "Effects of vowel nasalization on the perception of vowel height", in *Nasalfest: Papers from a Symposium on Nasals and Nasalization*, edited by C. A. Ferguson, L. M. Hyman, and J. J. Ohala, 373–388 (Language Universals Project, Stanford University, Stanford, Calif.).

Wright, J. T. (**1986**). "The behavior of nasalized vowels in the perceptual vowel space", in *Experimental Phonology*, edited by J. J. Ohala and J. J. Jaeger, 45–67 (Academic, Orlando, Fla.).

A Word lists

The three word lists used in this study are presented below. Within each list, the words are grouped into the "minimal sets" out of which the list was originally constructed: minimal pairs {CV(C), C$\tilde{\text{V}}$N} and minimal triples {CV(C), C$\tilde{\text{V}}$N, C$\tilde{\text{V}}$(C)} (Bengali only). For each word, the list shows the orthographic form (except in Bengali which is not written in a Roman script), the pronunciation in IPA, and the gloss (for languages besides English).

A.1 English word list

bob	/bab/	hut	/hʌt/	dib	/dɪb/
bomb	/bam/	hun	/hʌn/	dim	/dɪm/
god	/gad/	sub	/sʌb/	dig	/dɪg/
gone	/gan/	sum	/sʌm/	ding	/dɪŋ/
bog	/bag/	bode	/boʊd/	kid	/kɪd/
bong	/baŋ/	bone	/boʊn/	kin	/kɪn/
bad	/bæd/	robe	/roʊb/	wig	/wɪg/
ban	/bæn/	roam	/roʊm/	wing	/wɪŋ/
bag	/bæg/	code	/koʊd/	dead	/dɛd/
bang	/bæŋ/	cone	/koʊn/	den	/dɛn/
dab	/dæb/	dude	/dud/	Jeb	/dʒɛb/
dam	dæm/	dune	/dun/	gem	/dʒɛm/
hag	/hæg/	sued	/sud/	head	/hɛd/
hang	/hæŋ/	soon	/sun/	hen	/hɛn/
lab	/læb/	tube	/tub/	bade	/beɪd/
lamb	/læm/	tomb	/tum/	bane	/beɪm/
pad	/pæd/	hued	/hjud/	paid	/peɪd/
pan	/pæn/	hewn	/hjun/	pain	/peɪm/
laud	/lɔd/	deed	/did/	tape	/teɪp/
lawn	/lɔn/	dean	/din/	tame	/teɪm/
pawed	/pɔd/	bead	/bid/	side	/saɪd/
pawn	/pɔn/	bean	/bin/	sign	/saɪn/
dug	/dʌg/	beep	/bip/	died	/daɪd/
dung	/dʌŋ/	beam	/bim/	dine	/daɪn/
gut	/gʌt/	deep	/dip/	pout	/paʊt/
gun	/gʌn/	deem	/dim/	pound	/paʊnd/
hub	/hʌb/	jib	/dʒɪb/	gout	/gaʊt/
hum	/hʌm/	Jim	/dʒɪm/	gown	/gaʊn/
hug	/hʌg/	bid	/bɪd/	Lloyd	/lɔɪd/
hung	/hʌŋ/	bin	/bɪn/	loin	/lɔɪn/

A.2 Spanish word list

paz	/pas/	“peace”
pan	/pan/	“bread”
parcha	/partʃa/	
pancha	/pantʃa/	
plasta	/plasta/	
planta	/planta/	“plant”
sal	/sal/	“salt”
san	/san/	“without”
salto	/salto/	“I jump”
santo	/santo/	“holy”
tal	/tal/	“so/than”
tan	/tan/	“so/than”
tes	/tes/	“teas”
ten	/ten/	“have (imp.)”
tres	/tres/	“three”
tren	/tren/	“train”
resta	/resta/	“it remains”
renta	/renta/	“rent”
fuerte	/fwerte/	“loud/hard”
fuente	/fwente/	“source”
pista	/pista/	“race track”
pinta	/pinta/	“he/she paints”
col	/kol/	“cabbage”
con	/kon/	“with”
dos	/dos/	“two”
don	/don/	“sir”
por	/por/	“by”
pon	/pon/	“put (imp.)”
sol	/sol/	“sun”
son	/son/	“they are”

A.3 Bengali word list

A.3.1 Pairs

/ʃik/	“bar/grid”	/tʰak/	“stacks”
/ʃiŋ/	“horn”	/tʰan/	“widow’s sari”
/piʈʰ/	“back”	/pɔtʰ/	“path”
/pin/	“pin”	/pɔn/	“promise”
/ʈip/	“forehead dot”	/dʒɔl/	“water”
/ʈin/	“tin”	/dʒɔŋ/	“rust”
/til/	“sesame seed”	/ʃɔkʰ/	“hobby/desire”
/tin/	“three”	/ʃɔŋ/	“clown”
/peʈ/	“stomach”	/lobʰ/	“greed”
/pen/	“pen”	/lom/	“body hair”
/tʃek/	“check”	/tʃul/	“hair”
/tʃen/	“chain”	/tʃun/	“lime”
/pʰel/	“throw”	/guʈ/	“molasses”
/pʰen/	“froth (boiled rice)”	/gun/	“qualities”
/tal/	“turbid water”	/gʰur/	“roundabout”
/tan/	“pretense”	/gʰun/	“termite”

A.3.2 Triples

/gʰaʈ/	“riverbank”	/kʰaʈ/	“bed”
/gʰam/	“sweat”	/kʰam/	“envelope”
/gʰãʈ/	“trick”	/kʰɔ̃dʒ/	“crease”
/kadʒ/	“work”	/tap/	“heat”
/kan/	“ear”	/gan/	“song”
/kãdʰ/	“shoulder”	/tãt/	“loom”
/bat/	“gout”	/ɖal/	“lentils”
/ban/	“flood”	/ɖan/	“right”
/bãdʰ/	“dam”	/ɖãʈʰ/	“arrogant”
/paʈʰ/	“jute”	/daʃ/	“servant”
/pan/	“betel leaf”	/dan/	“gift”
/pãtʃʰ/	“five”	/dãt/	“teeth”
/bʰat/	“cooked rice”	/dʰaʈ/	“person’s nature”
/bʰan/	“pretense”	/dʰan/	“grain”
/bʰãɽ/	“clay cup”	/dʰãtʃ/	“style”

(continued)

/ʃat/	“seven”	/gɔt/	“tune”
/ʃan/	“whetstone”	/gɔm/	“wheat”
/ʃãʃ/	“kernel”	/gɔ̃d/	“to smell”
/tʃal/	“rice”	/k^{h}ur/	“hoof”
/tʃan/	“bathe”	/k^{h} un/	“murder”
/tʃãd/	“moon”	/k^{h}ũt/	“fault”
/rat/	“night”	/dip/	“island”
/ran/	“ran (in cricket)”	/dik/	“direction”
/rãd^{h}/	“to cook”	/din/	“day”
/hat/	“hand”	/tʃil/	“kite (bird)”
/ham/	“measles”	/tʃit/	“supine”
/hãp/	“gasp”	/tʃin/	“China”
/ak^{h}/	“sugarcane”	/dʒor/	“force”
/am/	“mango”	/dʒom/	“god of death”
/ãʃ/	“scale”	/dʒõk/	“leech”
/bæg/	“bag”	/kol/	“lap”
/bæŋ/	“frog”	/kon/	“corner”
/bãk/	“bend (n.)”	/kõtʃ/	“clothtuck”
/gatʃ^{h}/	“tree”	/go/	“cow”
/gan/	“song”	/gol/	“round”
/gãʈ/	“knot”	/gon/	“count”
/gã/	“village”	/gõp^{h}/	“mustache”
/ges/	“gas”		
/gen/	“knowledge”		
/gæ̃t/	“strong posture”		