Acoustic analysis and synthesis of nasal codas in Mandarin Chinese

by

Chi-Yu Liang

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 10, 2002

Author . . . .
Department of Electrical Engineering and Computer Science
May 10, 2002

Certified by . . . .
Kenneth N. Stevens
Professor, Research Laboratory for Electronics
Thesis Supervisor

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

Acoustic analysis and synthesis of nasal codas in Mandarin Chinese

by

Chi-Yu Liang

Submitted to the Department of Electrical Engineering and Computer Science
on May 10, 2002, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The aim of this study is to identify acoustic features of the two nasal codas, [n] and [ŋ], that exist in Mandarin Chinese. Special attention was placed on the synthesis of nasal syllables with differing numbers of pole-zero pairs. First, appropriately selected nasal utterances were copy-synthesized from the spoken utterances of a native speaker of the language. The nasal utterances were copy-synthesized with zero, one, and two nasal pole-zero pairs. A perception test was administered to provide a measure of the importance of nasal pole-zero pairs in the perception of nasals in Mandarin Chinese. Upon analysis of the collected data, it was concluded that each additional pole-zero pair improved both the perceptual quality and the accuracy of identification for the two nasal codas. Also, the high-frequency pole-zero pair provided more increase in perceptual quality in the velar nasal than for the alveolar nasal, providing a possible cue for the velar/alveolar distinction. These results should be taken with the understanding that there were potentially significant sources of error in this study. Characterizing the perceptual distinction between zero, one, or two nasal pole-zero pairs in Mandarin nasal codas will help improve the effectiveness of Chinese speech recognition and synthesis systems.

Thesis Supervisor: Kenneth N. Stevens
Title: Professor, Research Laboratory for Electronics

Acknowledgments

This research was conducted at the Speech Communication Group of the Research Laboratory for Electronics at M.I.T.

My most heartfelt thanks to Kenneth N. Stevens for his guidance and patience throughout the duration of this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This is a study of the analysis, synthesis, and perception of nasal codas in Mandarin Chinese. The two nasal codas in this language are [n] and [ŋ]. In order to study these nasal codas, six carefully chosen utterances were recorded by a native speaker of Mandarin Chinese. The nasal utterances were then analyzed and digitally synthesized to mimic the original utterance. Existing acoustic theory of nasals involves the introduction of pole-zero pairs into the transfer function of a nasal utterance. To study the effect of these poles and zeroes on the perception of these nasal utterances, each utterance was copy-synthesized with two pole-zero pairs, then with one pole-zero pair, and finally with no pole-zero pairs. Synthesized and original utterances were included in a perception test which was administered to four native speakers of Mandarin Chinese. The analyzed data show that there is a direct relationship between the number of pole-zero pairs and the perceptual quality of the synthesized utterances as well as their incidence of correct identification. Additionally, they suggest that the effects of the high-frequency pole-zero pair could be a distinguishing acoustic feature between alveolar and velar nasal codas. These data, however, were subject to many sources of error. There were additional conclusions on the identification of these nasal codas as well as the effects of the carrier phrase on their perception.

The remainder of this chapter will explain in more detail the motivation and background surrounding research of nasal utterances.

Chapter two describes the methodology of spectral analysis and copy-synthesis of the nasal utterances. It follows with a description of the development and administration of the perception test.

Chapter three describes the analysis of data and the results and trends of the perceptual test.

Chapter four describes the conclusions that can be drawn from the data as well as other research which was conducted in the development of this study. Additional observations on nasal codas in Mandarin Chinese are also explored in this chapter.

1.1 Motivations

Only two nasal codas exist in Mandarin Chinese: [n] and [ŋ] [3]. In fact, with the exception of the syllable [ɛr], there are no other syllables in Mandarin Chinese which end in consonants which are not nasal. In other words, except for [ɛr], the two nasal codas [n] and [ŋ] are the only non-vowel codas in Mandarin Chinese [9]. In fact, of the monosyllabic words in the language, over 40% of them have a vowel followed by a nasal coda [1]. The distinction between the two nasal codas is particularly interesting not only because of their frequency in the language, but because it is often difficult for non-native speakers of Mandarin as well as speakers of different Chinese dialects to distinguish between the two endings [4]. Thus the study of nasal codas in Mandarin Chinese is essential for the development of Chinese speech synthesis and recognition systems. Additionally, although there has been significant research on nasals in English, there are few studies of nasals in Mandarin Chinese. This study hopes to more fully characterize the acoustically significant properties of nasal codas in Mandarin.

1.2 Physiology

A nasal consonant is produced by a closure within the oral cavity and a velopharyngeal opening [19]. Both of the nasal codas in Mandarin Chinese require a partial or full closure within the oral cavity as well as an open velopharyngeal port. The two nasal codas differ physiologically only in their place of articulation: [n] is created with an alveolar constriction while the constriction for [ŋ] is velar. As the oral closure is made, the velopharyngeal port opens, creating a new acoustic path and a new path for airflow. In a vowel-nasal (V-N) syllable, the velopharyngeal port starts to open before the oral closure is complete, which results in nasalization of the pre-nasal vowel. When the closure in the oral cavity becomes complete and the vocal folds continue to vibrate, there is acoustic transmission through the nasal passage, and nasalization in the form of a nasal murmur is heard. Although the physiological changes are gradual, the introduction of the nasal cavity and the oral closure in the vocal tract configuration for the V-N syllable usually result in an abrupt change in the spectrogram at the V-N boundary [7]. The effect of nasalization is to change the formant frequencies as well as to introduce pole-zero pairs into the transfer function for the nasal utterance. The theory behind these spectral changes is based on both acoustic properties and experimental evidence.

1.3 Acoustics

The acoustic tube model for the vocal tract in the case of a non-nasal vowel is of a simple tube with a back section depicting the pharyngeal area and a front section depicting the oral cavity. The width of the front and back tubes may be adjusted to model different front, back, high and low vowels. This acoustic system has natural frequencies which correspond to the formant frequencies (the poles of the volume-velocity transfer function) of speech. The cross-sectional areas of the front and back tubes as plotted against distance from the glottis is known as the area function. As the tongue moves from its configuration from the vowel to the velar or alveolar constriction, the area functions of the front and back tubes may change, resulting in the shifting of the natural frequencies of the system.

Additionally, when the velopharyngeal port begins to open, there is a side-branch tube representing the nasal cavity added to the model of the pre-nasal vowel. While the velopharyngeal port is open but before the oral closure is made complete, there is acoustic output through both the mouth and the nose, resulting in a nasalized vowel. During this vowel nasalization there is nasal coupling and the introduction of pole-zero pairs to the transfer function. Once the oral closure is complete, the acoustic output is only from the nose, and the nasal murmur is heard. This causes the natural frequencies of the transfer function to change and the nasal pole-zero pairs to separate and shift.

The acoustic passage through the nose also introduces more acoustic loss in the nasal cavity. This results in higher bandwidths of the resonant frequencies, particularly of F_1.

1.3.1 Formant Transitions

The formant transitions seen in a V-N syllable can vary depending on both the place of articulation for the nasal and on the particular pre-nasal vowel. The transitions are easily explained through the acoustic tube model of the vocal tract. Before the V-N transition in the syllable, the vocal tract configuration is an open tube with the appropriate widths for the back and front sections corresponding to the physiology of the pre-nasal vowel. As the alveolar or velar constriction begins to be made with the tongue, the open tube begins to gradually perturb until a complete closure is made. The different perturbations that occur for alveolar and velar consonants result in different formant transitions based on their acoustic models. These formant transitions are similar to those of stop consonants with the same place of articulation [19].

Alveolar Nasal, [n]

For the alveolar nasal consonant, the oral closure is made with the tongue blade at a distance of 5-6 cm from the velopharyngeal port, or about 1.5 to 2.5 cm from the lips. In order to make the alveolar constriction, the tongue body must move to a more fronted position and the tongue blade must rise towards the hard palate. Thus if the preceding vowel is a front vowel, there will be little movement of the tongue body when transitioning to the nasal. In contrast, if the preceding vowel is a back vowel, the tongue body must move from a backed position to a more fronted one, causing F_2 to rise. Similarly, for a high vowel, there will be little transitional movement from the vowel to the nasal and thus little change in the frequency of F_1. For a low vowel, there will be a prominent decrease in F_1 because of the movement of the tongue from a low to a high position [19].

Velar Nasal, [ŋ]

The velar nasal consonant [ŋ] is produced with a closure of the tongue body at the soft palate at a distance of 5-7 cm from the lips. Because the closure is made with the tongue body, the rate of decrease of the area function of the oral constriction is much slower for velar consonants than for alveolar consonants. The slower speed at which the oral constriction is made results in a longer time in which the nasal cavity and the anterior oral cavity are both excited, resulting in a more nasalized vowel.

For the velar closure, the lowest natural frequency of the cavity anterior to the closure is relatively close to a resonant frequency of the back cavity. Thus as the V-N transition is made, we should expect F_2 and F_3 to move closer together. The frequency of this convergence varies depending on the preceding vowel. For a front vowel, the tongue body is in a more fronted position, resulting in a smaller front cavity and thus a higher front cavity resonance. Similarly for back vowels, the front cavity is slightly larger, resulting in a lower front cavity resonance. Thus the convergent frequency of F_2 and F_3 for the velar nasal is higher when it is preceded by a front vowel than when it is preceded by a back vowel. The analysis for F_1 remains the same for [ŋ] as it was for [n]. There will be more transitional movement of F_1 near the V-N boundary for a low vowel than for a high vowel [19].

1.3.2 Poles and Zeroes

As the vowel transitions into the nasal coda, a low-frequency pole-zero pair, previously undetectable in the transfer function, begins to increase in frequency. As the pole and zero increase in frequency during the V-N transition, they also separate from one another. Once the pole-zero pair begin to separate, they no longer cancel each other out and therefore begin to have an effect on the transfer function. When the oral closure is made complete, there is an abrupt change in the acoustic model; only the nasal cavity and the cavity posterior to the oral constriction are excited by the glottal source. Although the cavity anterior to the oral constriction is still a part of the physical acoustic system, it is no longer excited. Thus during the nasal murmur, the zero in a pole-zero pair must move to a frequency to obscure the lowest resonant frequency of the front cavity. This zero is at the frequency at which the impedance looking into the oral cavity from the velopharyngeal opening is zero. This low-frequency zero is often close to F_2, the second natural frequency of the vocal tract. The corresponding nasal pole is usually around 1 kHz. Similarly, there is a high-frequency pole-zero pair which is in the 2 kHz range during the vowel. At the V-N boundary the zero quickly jumps to obscure poles in the region of F_4 and F_5 while exposing the nasal pole in the F_2-F_3 range [19].

The frequencies of the poles and zeroes can vary with each individual speakers' nasal coupling. Despite these variations, there are documented typical values for the frequencies of nasal poles and zeroes for [n] and [ŋ] [2]. The values for the high-frequency pole-zero pair tend to vary a lot more than those for the low-frequency pole-zero pair.

Alveolar Nasal, [n]

The primary resonance in the spectrum of a nasal consonant is a Helmholtz resonance due to the compliance of the vocal tract and the acoustic mass of the nasal passage. It occurs at about 250 Hz. The low-frequency pole-zero pair becomes separate at the V-N boundary and the pole rises to about 1 kHz, creating a small peak along the edge of F_1, while the zero rises to 1400-1900 Hz to obscure F_2, the resonance of the oral cavity anterior to the closure. The high-frequency pole-zero pair originates at approximately 2 kHz. During the nasal murmur, however, the zero rapidly increases to obscure the resonances around 4500 Hz, leaving the pole to be exposed in the 2000 Hz range [19].

Velar Nasal, [ŋ]

The primary resonance of the spectrum in the nasal consonant is again at 250 Hz, followed by the secondary nasal peak at 1000-1200 Hz due to the low frequency pole. The low frequency zero is in the range of 1800-3000 Hz to obscure the resonance of the oral cavity anterior to the velar closure. The second pole-zero pair is at higher frequency, with the pole at about 2100 Hz and the zero at 4 kHz or higher [19].

1.4 Existing Research

A wealth of research in the area of nasal consonants and nasalized vowels has been conducted. Much of the research which concerns analysis and synthesis of nasals, however, has been focused on nasals in English and French. There are limited studies written in English which focus on the acoustic properties of nasals in Mandarin Chinese.

1.4.1 Research in English

Much of the established research on the English nasals [m], [n] and [ŋ] focused on the importance of formant transitions and/or the nasal murmur as cues [12] [11] [17] [13]. It has been well established that subjects can repeatedly correctly identify nasals which have been synthesized with a "neutral" murmur and varying formant transitions [11]. The transitions associated with with labial, alveolar and velar nasals are the same transitions which have been well documented in the cases of labial, alveolar and velar stop consonants. This research is consistent with the acoustic theory presented in the previous section.

In separate studies conducted by Malecot [14] and Recasens [15], it was confirmed that the nasal murmurs are not completely neutral across English nasals, but that their differences did not contribute as significantly to perceptual distinctions as the formant transitions did. However, studies have also shown that there are perceptual cues in the nasal murmur that are consistent with differing places of articulation [11].

Thus both the nasal murmur and the formant transitions are thought to play roles as perceptual cues. Furthermore, Kurowski and Blumstein [11] conducted a study which supported the view that the cues contained in both the nasal murmur and the formant transitions form an "integrated property" by which to identify nasals. They assert that neither the nasal murmur nor the formant transitions are, on their own, a necessary and sufficient perceptual cue for place of articulation.

Another focus of study on English nasals is spectral characterization. Metrics of spectral characterization have been developed at least two ways. First, a metric was developed by measuring the change in energy in certain frequency bands in the transition from nasal to vowel release [12]. Based on the location of the zeroes (antiformants) in the spectrum, labial nasals should have a greater change in energy than alveolar nasals in the 396-770 Hz range than in the 1265-2310 Hz range. This metric achieved at best a 89% correct identification score. Another proposed metric was similar to the first except that it was based on relative frequencies in the N-V boundary [17]. The "difference spectrum" was calculated by subtracting the vowel spectrum from the murmur spectrum. By noting the maximum and minimum difference spectrum values, a 77% correct identification score was achieved.

Although these metrics did a reasonable job of identifying different nasals, the tests used on these metrics were only conducted with nasal-initial syllables. In fact, it has been shown that there is a less abrupt spectral change in V-N syllables than in N-V syllables due to the nasalization of the vowel that occurs in the V-N case [19] [16]. There is thus reason to believe that these metrics would not be as accurate in identifying nasals in a V-N syllable [6], such as the ones in Mandarin Chinese.

1.4.2 Research in Chinese

Significant research has been done in the synthesis of Mandarin Chinese, especially in the areas of perfecting the synthesis and analysis of tones [8] [20] [18]. However, there has not been very much research published in English regarding the synthesis of the two nasal codas in the language. One significant study by M.Y. Chen [2], was focused on determining the acoustic cues for nasal codas in Mandarin Chinese. Chen's study was much more extensive and complex than the one conducted here. She used combinations of two-word phrases which included nasal codas or nasal initials. In her study, she focused not on the acoustic properties of the two nasal codas, but instead focused on detecting a vowel-nasal boundary in Mandarin Chinese. She concluded that the distinction between the two nasal codas may be made by the formant frequencies of the preceding vowel as well as by analyzing the vowel-nasal boundary. Based on acoustic models of speech production, a vowel-nasal boundary should be distinguished by an abrupt change in the amplitudes of the formant prominences. Chen found that when a complete oral closure is made, the amplitude of the first four formants is at least 11 dB higher for the [n] coda than for the [ŋ] coda.

Thus far, there have been no studies of the role of the number of pole-zero pairs as possible cues for nasal distinctions.

Chapter 2

Methodology

The aim of this study is to characterize the perceptual significance of nasal pole-zero pairs of nasal codas in Mandarin Chinese. In order to do this, selected syllables with the syllable-finals [n] and [ŋ] were copy-synthesized from utterances of a native speaker of Mandarin. Each syllable was synthesized first with two pole-zero pairs, then with only one pole-zero pair and finally with no pole-zero pairs. The effect of the number of pole-zero pairs in the perception of the two nasal codas [n] and [ŋ] was measured in perception tests of native speakers of Mandarin Chinese. The results of this research are intended to strengthen understanding about the acoustic attributes and distinguishing features of nasal codas in Mandarin Chinese, particularly for speech synthesis and recognition systems.

2.1 Utterance Selection

There are five pairs of word-finals which exist in Mandarin Chinese as [n] - [ŋ] pairs: "-an" and "-ang," "-en" and "-eng," "-ian" and "-iang," "-in" and "-ing," and "-uan" and "-uang." The two pairs that include a diphthong, however, are not easily confused to listeners of Mandarin Chinese because the diphthong causes the elongated vowel in the two [n] and [ŋ] finals to shift. For instance, "-ian" and "-iang" are pronounced [iεn] and [iɑŋ], respectively. The front vowel [ε] is easily distinguished from the back vowel [ɑ], and thus there is relatively little perceptual confusion between the two nasal codas. Similarly, "-uan" and "-uang" are pronounced [uɛn] and [uɑŋ], respectively. Thus only the three remaining pairs of word-finals are more difficult to distinguish and are the focus for this study.

Although there are only three remaining pairs, the four tones in Mandarin as well as the many homophones in the language give rise to several dozens of pairs of words which may often be misheard and misunderstood. It is thus important for a listener to be able to distinguish between the two nasal codas in order to comprehend spoken Mandarin Chinese and also for researchers to understand the acoustic characteristics of nasals for speech applications.

For every chosen nasal syllable with the coda [n], the corresponding nasal syllable ending in [ŋ] is also chosen, forming a pair. These pairs were constructed to consist of words of the same tone that exist in Mandarin Chinese. Tone three was avoided so as to eliminate the added complication of possible glottalization. Nasal syllables were chosen to be without a nasal initial and preferably with a stop consonant initial (for ease of synthesis). A nasal initial would cause the following vowel to be nasalized, thus making it harder to discern the effect of the nasal final. Three pairs of words with nasal codas are chosen to be representative of the three distinct pairs of nasal codas described above. The three pairs, all with consonant-vowel-nasal (CVN) structure, are pín and píng, bēn and bēng, and tán and táng. The accents indicate the tones.

2.2 Recordings

A native male speaker of Mandarin Chinese was chosen to record utterances for copy-synthesis. The utterances were comprised of the nasal syllables listed above, each inserted as the CVN in the carrier phrase "shūo CVN bȧ," which may be translated as "say CVN now" in English. The sequence of six utterances was recorded three times in succession using an Electro-Voice D054 microphone which was placed approximately six inches from the speaker's mouth. The utterances were recorded with a Nakamichi LX-5 discrete-head cassette deck onto an analog cassette tape in a quiet noise-reduced room. Once the utterances were on the cassette tape, they were digitized at a 10 kHz sampling rate using the RECORD program developed by D.H. Klatt. RECORD is able to digitize recordings on a cassette tape at any specified sampling rate using one of a few available low pass anti-aliasing filters. The digitized sequence of utterances was then spliced and saved as individual utterances within the RECORD program. Each individual utterance consisting of the nasal syllable and carrier phrase was then opened in the RECORD program in order to excerpt the nasal syllable from the rest of the utterance. After the nasal syllable was synthesized, it was re-inserted into the carrier phrase using the CONCAT program, which can concatenate a series of digital speech utterances.

2.3 Analysis

Before synthesis of a nasal syllable may begin, analysis was conducted on the original nasal syllable. The LSPECTO program was run to generate a spectrogram, formant tracks, a fundamental frequency track, and a list of estimated formant and fundamental frequency values throughout the duration of the utterance. The spectrogram, formant tracks, and F_0 track of the nasal syllable tán are shown in Figures 2-1, 2-2 and 2-3, respectively. The formant and fundamental frequency tracks are the best estimates of the LSPECTO program, and were used with the spectrogram as rough guidelines in the synthesis process.

The spectrogram shows the burst for the /t/ and the vowel transition into the nasal at approximately 210 ms. The discontinuity here is representative of the type of discontinuity that is present in all of the nasals in this study. There are three visible changes in the spectrogram which are consistent with the acoustic theory described earlir in Section 1.3. First, F_1 decreases at the discontinuity. The decrease in F_1 is very noticeable in this case because the nasal coda in tán is preceded by the low vowel /ɑ/. In the case of the syllables pín and píng, however, there is not as noticeable of a decrease in F_1 because of the high vowel preceding the nasal coda. Second, there is significantly less high frequency energy after the discontinuity. This is in part due to the lowering of F_1 but is also due to the introduction of poles and zeroes into the

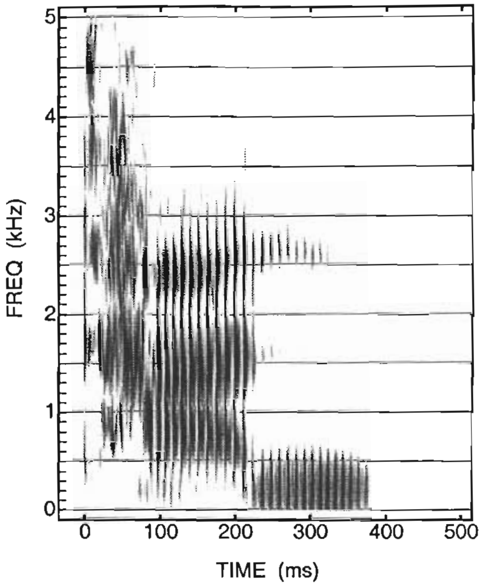

Figure 2-1: Spectrogram of tán

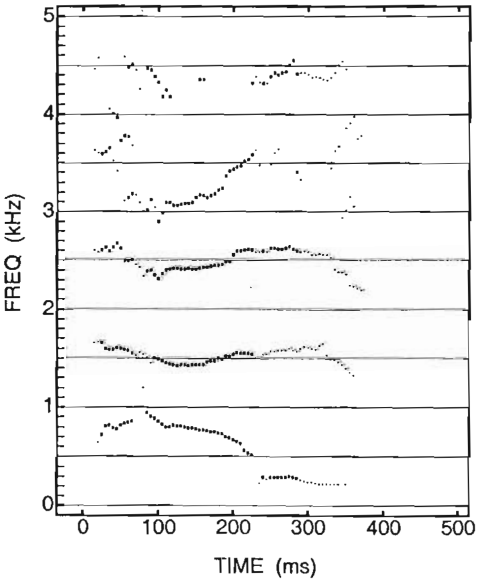

Figure 2-2: Formant frequency tracks of tán

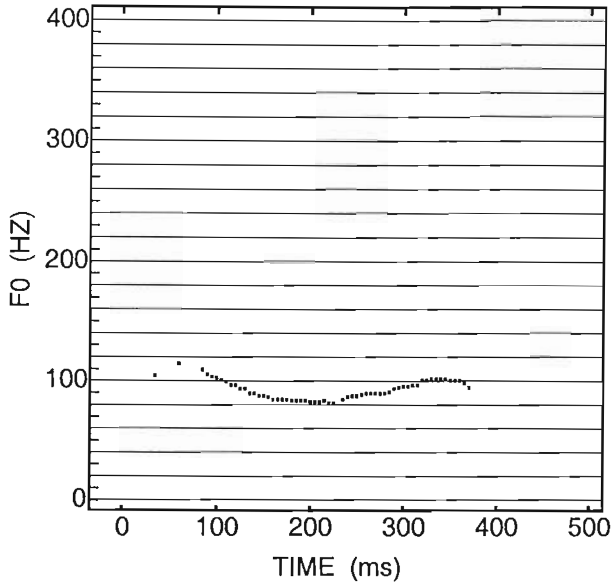

Figure 2-3: Fundamental frequency track of tán

transfer function. Third, the bandwidths of the formant frequencies, particularly F_1, increase because of the increased acoustic loss of the fleshy nasal cavity.

Upon closer inspection, the spectral differences before and after the discontinuity become apparent. Instead of the clear, successive formant frequencies that are seen in the vowel, only F_1 and what appears to be F_3 are clearly visible in the spectrogram after the discontinuity. This type of spectral change is due to the introduction of pole-zero pairs into the transfer function of the speech. The low-frequency pole is of very high bandwidth at approximately 850 Hz while the low-frequency zero rises to obscure F_2 at about 1500 Hz. The high-frequency pole rises to reinforce F_3 at about 2500 Hz while the high-frequency zero rises to obscure most of the higher frequency components at 4500 Hz. These frequencies are consistent with the theory presented in Section 1.3.2.

2.4 Synthesis

2.4.1 Synthesizer Basics

Once this initial analysis of the original utterance was complete, synthesis was begun. Syntheses of the nasal syllables were conducted using the KLSYN88 synthesizer, developed by D.H. Klatt [10]. The synthesizer is imbedded in a UNIX program called XKL, which is capable of a variety of types of spectral analyses of digital utterances. KLSYN88 is a formant synthesizer which uses cascaded and parallel source-filter pathways in order to generate a speech output. With the correct manipulations, it is capable of synthesizing both male and female voices such that a listener would not be able to distinguish between an originally produced sound and its copy-synthesized version [5]. It is an effective synthesizer because of its various control parameters, most of which may be time-varying, allowing for careful manipulation of speech spectra. Although only the most relevant parameters are discussed here, a full listing and description of all KLSYN88 control parameters is provided in Appendix A, Table A.3. The parameters that were used most often in these syntheses of nasal syllables are listed in Table 2.1. All of the parameters listed in this table except for DU and SR can be time-varying. Time-varying parameters were entered in to KLSYN88 in a piecewise-linear fashion.

2.4.2 Synthesis of Vowel

The method of synthesis used here consisted of a series of sequential steps and manipulations, followed by repeated trial-and-error tweaking of parameters in order to copy-synthesize the original utterance as closely as possible. First, the basic parameters were set to their correct values. The synthetic sampling rate, SR, was set to 10 KHz, as this was the sampling rate of the original digital speech. The duration of the synthetic utterance, DU, was set to the appropriate length equal to the length of the original nasal syllable. The durations of the nasal syllables ranged from 300 ms to 400 ms. Next, the basic time-varying parameters were added. The fundamental

Table 2.1: Some KLSYN88 Parameters

SR	sampling rate of the digitized speech
DU	duration of speech
F0, F1, F2, F3, F4, F5	fundamental and formant frequencies
B1, B2, B3, B4, B5	bandwidths of formant frequencies
AH	amplitude of aspiration
AF	amplitude of frication
AV	amplitude of voicing
OQ	open quotient
TL	spectral tilt
FNZ, FNP	frequency of low-frequency pole-zero pair
FTZ, FTP	frequency of high-frequency of pole-zero pair.
BNZ, BNP, BTZ, BTP	bandwidths of the two pole-zero pairs.

frequency track and the list of estimated values of fundamental frequency over time were used as a rough estimate for the F_0 which was entered into KLSYN88. F_0 had a non-zero value only for the times in which voicing occured. Using the formant tracks and the list of estimated formant frequencies over the duration of the original utterance, the piecewise-linear plots of all five formant frequencies during the vowel were entered. However, the formant tracks were not always reliable, especially within the nasal portion of the utterance. For example, with the introduction of the low-frequency nasal pole-zero pair, F_2 became obscured while a new peak in the spectrum was seen at approximately 900 Hz. The locations of the formants during the V-N transition and during the nasal were therefore entered using an educated estimate based on the acoustic theory of alveolar and velar nasals and the vowel which preceded it. As described above, for alveolar nasals there were slight increases in the frequencies of F_2 and F_3 if they were preceded by a back vowel while for velar nasals, F_2 and F_3 tended to come together as the velar closure was made. The synthesized F_0 and formant contours for tán are shown in Figures 2-4 and 2-5, respectively.

After entering the formant frequencies, the next step in synthesis required adjusting the bandwidths of the formants. The Discrete Fourier Transform (DFT) spectrum of the original utterance and the thus-far synthesized utterance were displayed using XKL. The spectra were compared at the same time during each utterance, during

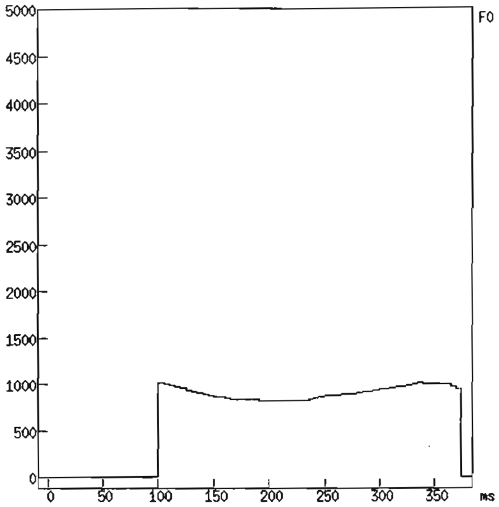

Figure 2-4: Fundamental frequency track of synthesized tán

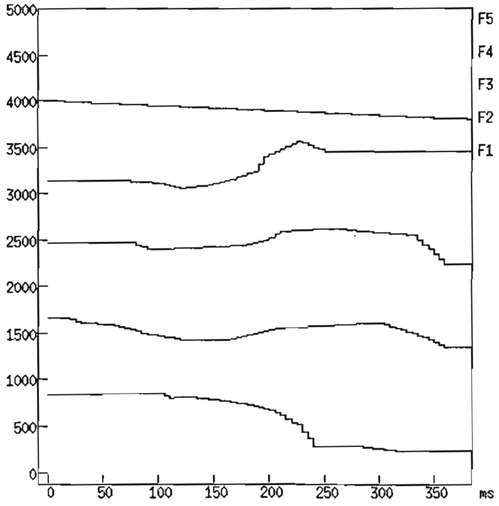

Figure 2-5: Formant frequency tracks of synthesized tán

a stable part of the vowel (before the V-N boundary), using a relatively large Hamming window (39.0 ms). These spectra provided good frequency resolution and were accurate in assessing the bandwidths of formant frequencies. By comparing the formant bandwidths of the original and synthesized utterance at the same points in time during the vowel, the proper adjustments were made to the control parameters B1, B2, B3, B4 and B5. The higher the bandwidth, the lower the amplitude of the formant peak. In general, doubling the bandwidth would decrease the amplitude of a formant peak by approximately 6 dB. The bandwidths of the synthesized utterance were adjusted so that the spread of a formant prominence was approximately across the same number of harmonics as in the original utterance. Thus far, the formant bandwidths were kept constant throughout the entire duration of the utterance.

Even if the bandwidths were adjusted so that the formant frequencies were of the correct width, they still might not have been of the appropriate height. Specifically, there were cases in these syntheses where there was too much high frequency energy. This was adjusted with spectral tilt, the TL parameter. The larger the tilt, the less high frequency energy. TL was adjusted to the correct value by comparing the DFT spectra of the synthesized and original utterance during the vowel. For all of the nasal utterances synthesized here, the maximum TL was 5, and it remained constant throughout the entirety of all the utterances.

Another important control parameter was the open quotient, OQ. These adjustments, similar to the ones made to the tilt, were made to decrease the differences between the original and synthesized utterance within the vowel of the nasal syllable. The open quotient is the percentage of time within a glottal cycle in which the glottis is open. The main effect of increasing OQ in this synthesizer is to increase the amplitude of the first harmonic in relation to the second. In speech production, an increase in OQ is normally accompanied by an increased breathiness in voice quality. For the particular male speaker recorded in this study, the open quotient was set as a constant between 43% and 50% in each synthesized utterance.

The amplitude of voicing, AV, is another parameter which is important in the quality of synthesized speech. In the synthesis of all six nasal syllables, it was shaped

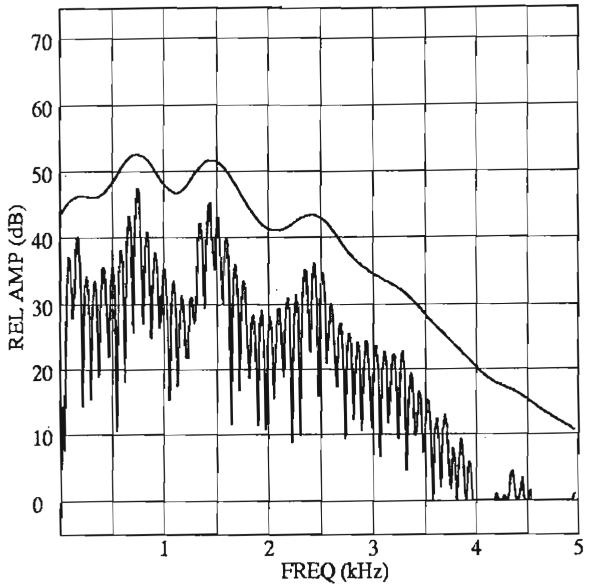

Figure 2-6: DFT Spectrum (Hamming window 39.0 ms long) of original tán in the vowel /ɑ/ at 175 ms

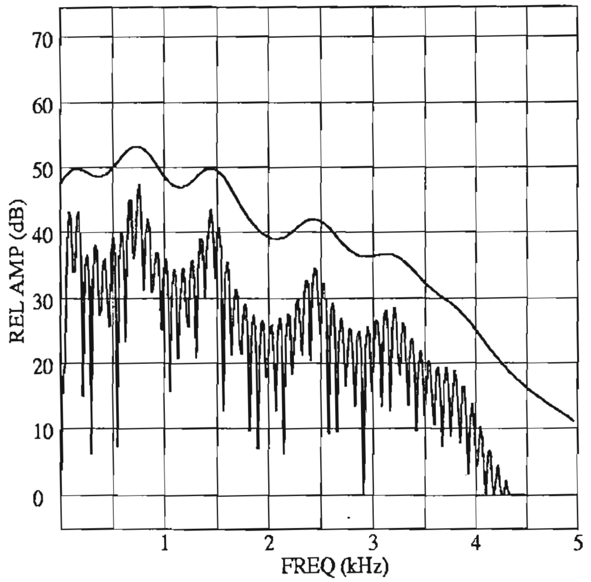

Figure 2-7: DFT Spectrum (Hamming window 39.0 ms long) of synthesized tán in the vowel /ɑ/ at 175 ms

as a time-varying parameter to account for the tapering of voicing that occurs at phrase beginnings and endings in regular speech. This provides a more natural contour to the amplitude of the synthesized speech. The typical time-varying AV ranged between 50 and 65 dB.

The last parameter that was adjusted while comparing the spectra during the vowel was the amplitude of aspiration, AH. In the cases of táng, tán, píng, and pín, there was significant aspiration noise from the voiceless stop consonant near the beginning of the vowel. To add noise to the vowel spectra of these nasal syllables, AH was time-varied from approximately 40 dB at the beginning of the vowel to 0 dB towards the middle of the vowel. There was little aspiration in the utterances of bēn and bēng, so lower-energy AH was incorporated into their syntheses.

Figures 2-6 and 2-7 show the DFT spectra and the smoothed spectra of the original and vowel-synthesized tán, respectively. The two spectra were both calculated with a relatively long Hamming window of 39.0 ms at a time of 175 ms, in the middle of the vowel /ɑ/. The synthesized formant frequencies, amplitudes and bandwidths were all similar in value to the original, as was the fundamental frequency.

After the synthesis steps described above, there were in some cases syntheses of the V-N syllable which already sounded similar to the original V-N utterance (as judged informally by the author). This was particularly true of all the utterances of pín. However, the spectrograms of all of the synthesized utterances, including those of pín, were visibly different from the spectrograms of the original utterances. The spectrograms for the synthesized and original utterances of pín are shown in Figures 2-8 and 2-9, and the synthesized and original utterances of bēng are shown in Figures 2-10 and 2-11. As can be seen from the spectrograms, there are still prominent second and fourth formants visible during the nasal murmur of the synthesized syllable for both utterances, while the originals have less energy in these frequency bands. This type of spectral difference made relatively less perceptual difference as compared to the original in the case of pín than for bēng.

All of the utterances, regardless of whether or not they sounded similar to the original, were modified still further in the synthesis process.

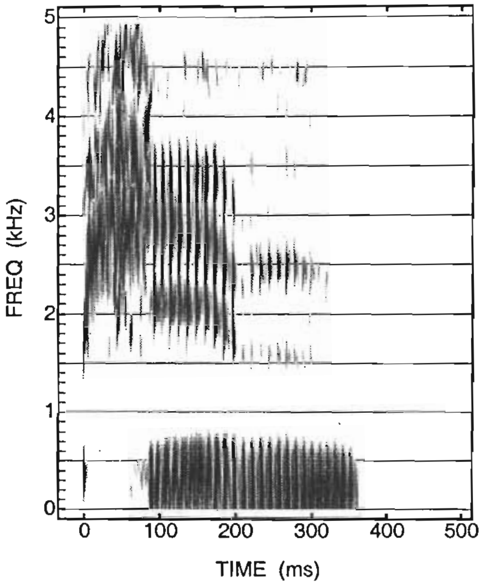

Figure 2-8: Spectrogram of original píŋ

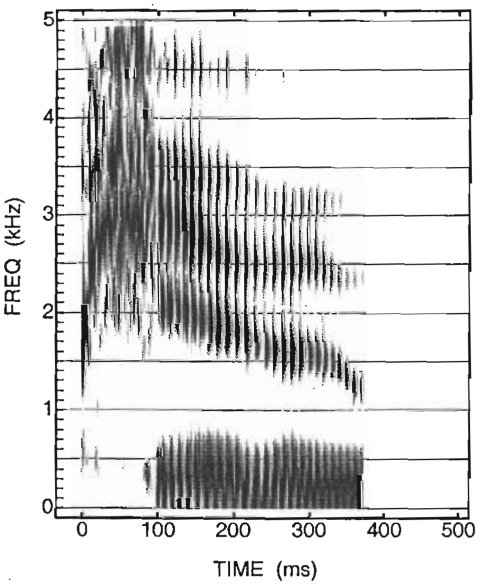

Figure 2-9: Spectrogram of vowel-synthesized píŋ

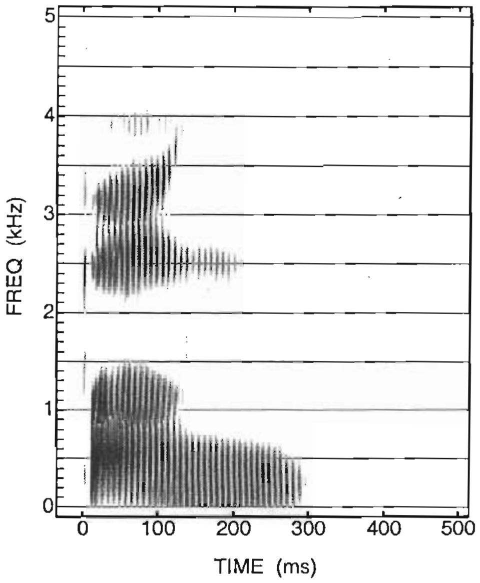

Figure 2-10: Spectrogram of original bēng

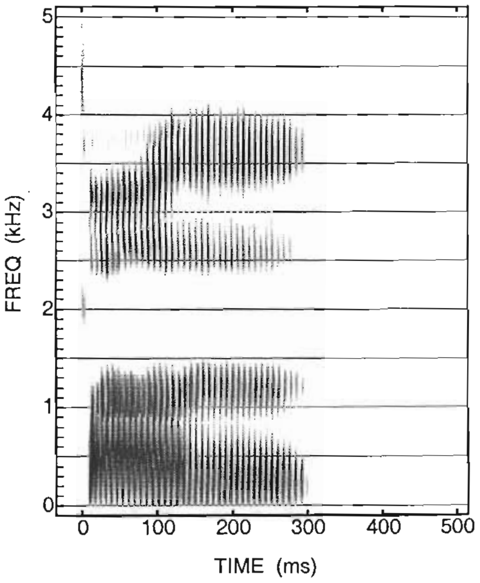

Figure 2-11: Spectrogram of vowel-synthesized bēng

2.4.3 Synthesis of V-N Boundary and Nasal Coda

Up until this point in the synthesis, only the properties of the vowel have been considered. The only special synthesis conducted to accommodate the V-N boundary and the nasal coda were adjustments of the formant transitions and formant frequencies into and during the nasal murmur. The remaining synthesis was focused on perfecting the V-N boundary and the nasal coda, not on the synthesis of the stop consonant preceding the vowel. The stop consonant burst was easily isolated from the original utterance using the RECORD program and then concatenated with the synthesized V-N syllable. Moreover, the stop consonant did not significantly affect the quality of the V-N boundary and the nasal coda, which was the focus of this study.

KLSYN88 allows manipulation of the spectra by modifying the frequency and bandwidth of two pole-zero pairs. Although acoustic theory can predict the general range for the frequencies and bandwidths of the poles and zeroes, determining the location of these pole-zero pairs was a somewhat subjective process. The pole-zero pairs were placed at the most appropriate frequencies based on both acoustic theory, spectrogram, and DFT spectra of the original utterance. With the acoustic theory in mind, the frequencies of the poles and zeroes were estimated by observing the changes in the formant prominences during the V-N boundary and into the nasal murmur. In accordance with the acoustic theory, the frequencies of the poles and zeroes were synthesized to fall within their theoretical ranges.

In the DFT spectra, we could see more clearly the effects of the pole-zero pairs as well as estimate their frequencies. The spectra in the nasal murmur of the original utterance of tán is shown in Figure 2-12. This spectrum was calculated with a Hamming window of length 39.0 ms at the time of 300 ms, which was in the middle of the nasal murmur. In contrast to the spectrum within the vowel /ɑ/ in Figure 2-6, this spectrum shows relatively little high-frequency energy but increased energy in the range of 100-400 Hz. That energy is the primary nasal resonance at around 200 Hz. There are also secondary resonances in the spectrum at approximately 1600 Hz and 2600 Hz. Although there is no peak at 1 kHz, the typical location of the low-frequency

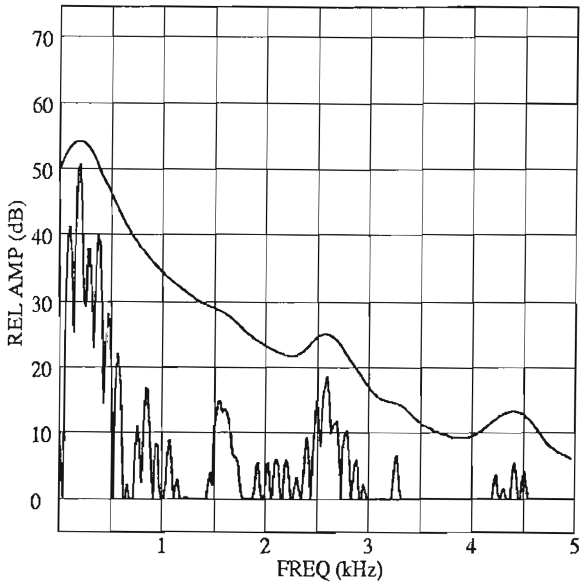

Figure 2-12: DFT Spectrum (Hamming window 39.0 ms long) of original tán within the nasal at 300 ms

nasal pole, there is a slight peak at approximately 800 Hz. Even though this value is low according to theory, it is still an acceptable value due to significant variation between speakers for the frequencies of pole-zero pairs. The spectrum also shows the frequencies of possible antiformants. The zeroes appear to be at approximately 1200 Hz and 3750 Hz.

The use of the DFT spectra throughout the nasal murmur allowed for careful placement of the poles and zeroes. In this particular case, the low-frequency zero was placed at 1300 Hz, the low-frequency pole at 850 Hz, the high-frequency zero at 3800 Hz, and the high-frequency pole at 2580 Hz. The synthesized frequencies for the low-frequency nasal pole-zero pair (FNP and FNZ) for tán are shown in Figure 2-13, and the synthesized high-frequency nasal pole-zero pair (FTP and FTZ) are shown in Figure 2-14.

The bandwidths of the nasal poles and zeroes were also adjusted to fit the original spectrograms as closely as possible. The process of adjusting the frequencies and bandwidths of the nasal poles and zeroes is one which involved significant trial and error. This process is made easier through the use of a short time-window for the

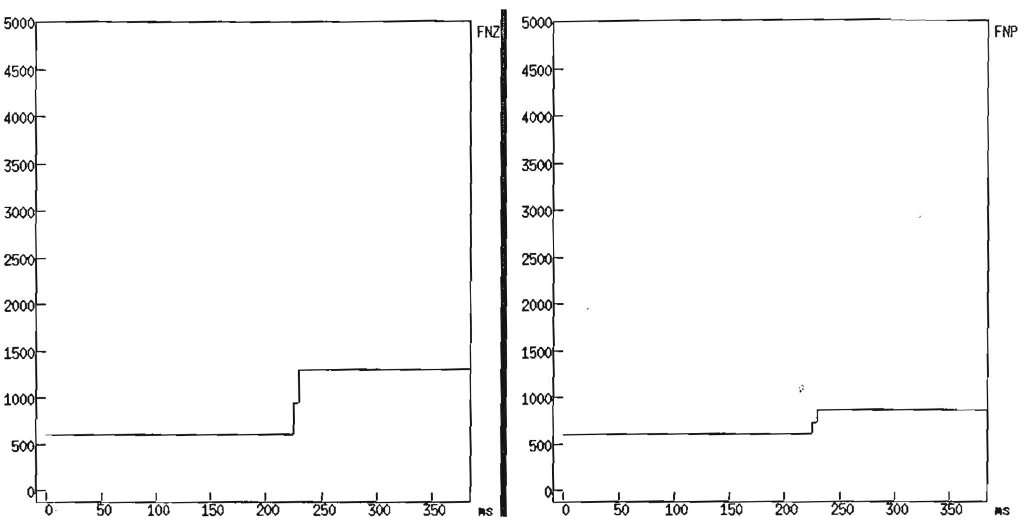

Figure 2-13: Synthesized frequencies of the low-frequency zero (FNZ) and pole (FNP) for tán

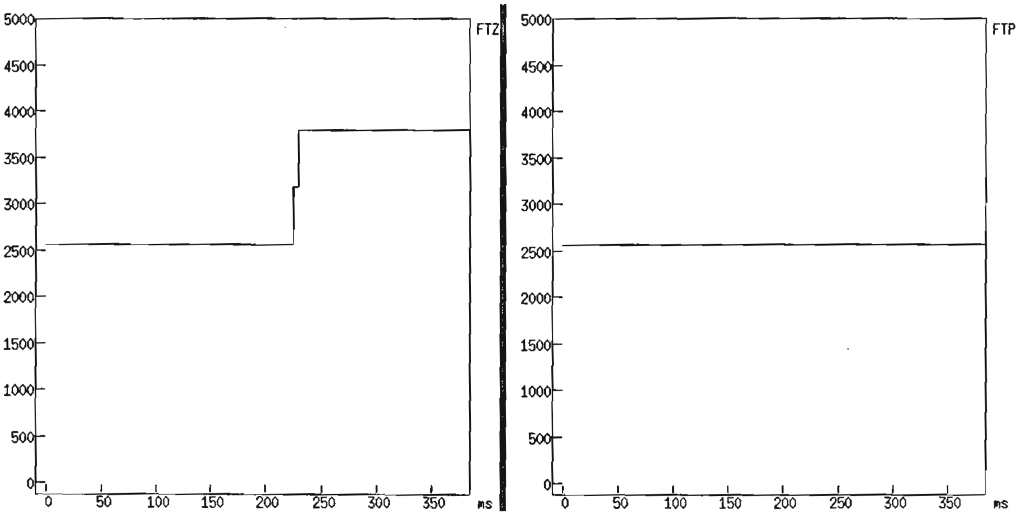

Figure 2-14: Synthesized frequencies of the high-frequency zero (FTZ) and pole (FTP) for tán

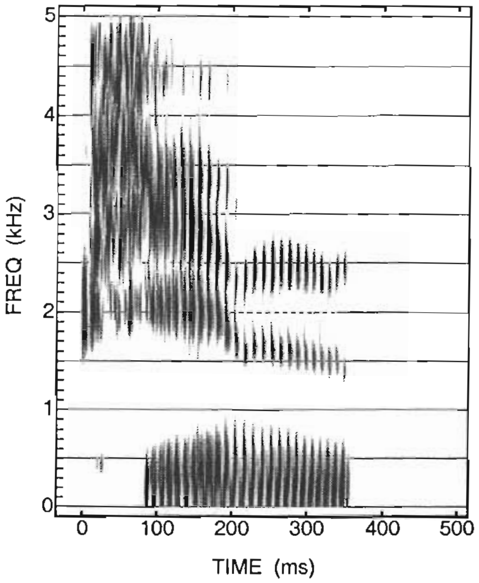

Figure 2-15: Spectrogram of synthesized pín with two pole-zero pairs

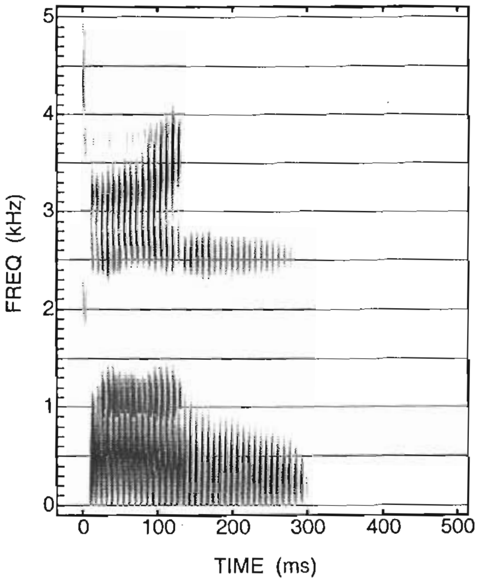

Figure 2-16: Spectrogram of synthesized bēng with two pole-zero pairs

DFT spectrum, where spectral variations were seen from pitch period to pitch period.

In the final stages of synthesis, small adjustments were also made to other control parameters. Formant bandwidths were changed from constant values to time-varying ones, which better matched the effect of the acoustic loss in the nasal cavity. Formant frequencies were also adjusted to better fit the spectra of the original utterances.

Once the utterance with two pole-zero pairs was successfully synthesized, the high-frequency pole-zero pair was removed to create the synthesis with one pole-zero pair. Similarly, the synthesis with no pole-zero pairs was created by removing both pole-zero pairs from the two pole-zero pair synthesis. The final results are utterances which both sound like and appear acoustically similar to the original utterances.

The new spectrograms, which include the effects of the two pole-zero pairs on the utterances pín and bēng, are shown in Figures 2-15 and 2-16, respectively. Compared to the vowel-synthesized versions (Figures 2-9 and 2-11), the spectra of pín and bēng with the two pole-zero pairs seem to be more similar to the spectra of the original speech (Figures 2-8 and 2-10). The full set of spectrograms for the utterance tán are also shown in Figures 2-17, 2-18, 2-19, and 2-20. These spectrograms show the original utterance as well as the three synthesized syllables with varying numbers of pole-zero pairs. These spectrograms show the effects of each pole-zero pair on the nasal murmur. With no pole-zero pairs, the second and fourth formant frequencies remained much more prominent in the nasal murmur than in the original tán murmur. Once the low-frequency pole-zero pair was added, however, F_2, F_3, and F_4 prominences decreased abruptly at the V-N boundary. This was due to the steep drop-off in amplitude of the pole at high frequencies. The synthesis which included the high-frequency pole-zero pair appeared to be most similar to the original. The high-frequency pole near F_3 resulted in an increased prominence of this formant, while the second and fourth formants had less energy.

These same effects of the two pole-zero pairs can be seen in the DFT spectra of the synthesized utterances of tán within the nasal murmur. With no pole-zero pairs, the spectrum within the nasal, shown in Figure 2-21, appeared similar to the spectrum of a vowel. There were no frequencies which appear to be near antiformants like the

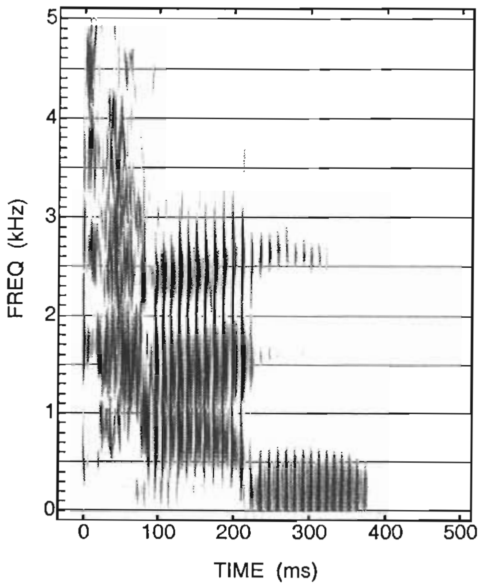

Figure 2-17: Spectrogram of original tán

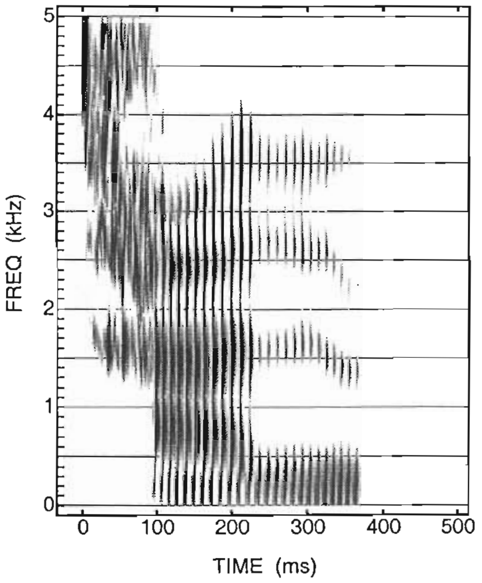

Figure 2-18: Spectrogram of synthesized tán with no pole-zero pairs

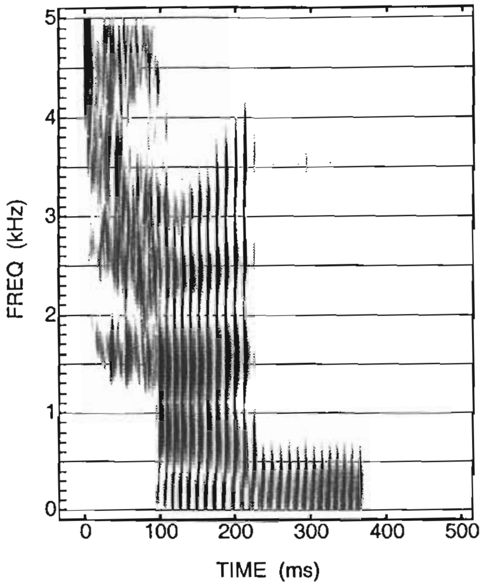

Figure 2-19: Spectrogram of synthesized tán with one pole-zero pair

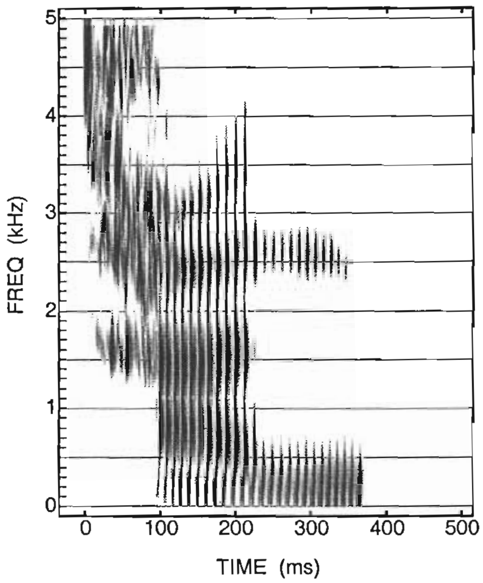

Figure 2-20: Spectrogram of synthesized tán with two pole-zero pairs

original spectrum in Figure 2-12. Also, the amplitudes of F_2, F_3, and F_4 did not match the orginal. With the addition of the low-frequency pole-zero pair in Figure 2-22, the low-frequency nasal pole-zero pair had a large effect on the spectrum. Formants at frequencies larger than that of the nasal zero became less prominent, but were still not of the correct relative amplitudes as they were in the original. With the addition of the high-frequency pole-zero pair, the spectrum in Figure 2-23 shows a more prominent F_3 as well as a slight prominence around 1600 Hz, both of which were consistent with the spectrum of the original utterance.

The variations in the spectrograms and DFT spectra for tán with each additional pole-zero pair are representative of the changes that were seen in the other five nasal utterances. Additional spectrograms and DFT spectra of the original and synthesized versions of the remaining five nasal syllables are included in Appendix B.

Although the manipulations in this study were of the number of nasal pole-zero pairs in the stimuli, the subjects of the perception tests heard only the synthesized audio output. Listeners therefore heard the spectral differences between the synthesized and original utterances, which masked the underlying variable of pole-zero pairs. Because the syntheses with two pole-zero pairs in most cases appeared spectrally to be most similar to the original, it was expected that these syntheses would also fare the best in the perception tests.

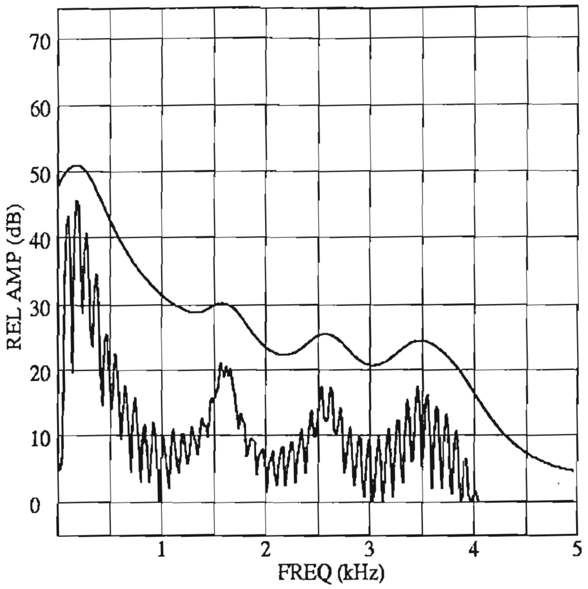

Figure 2-21: DFT Spectrum (Hamming window 39.0 ms long) within the nasal coda at 300 ms of the synthesized tán with no pole-zero pairs

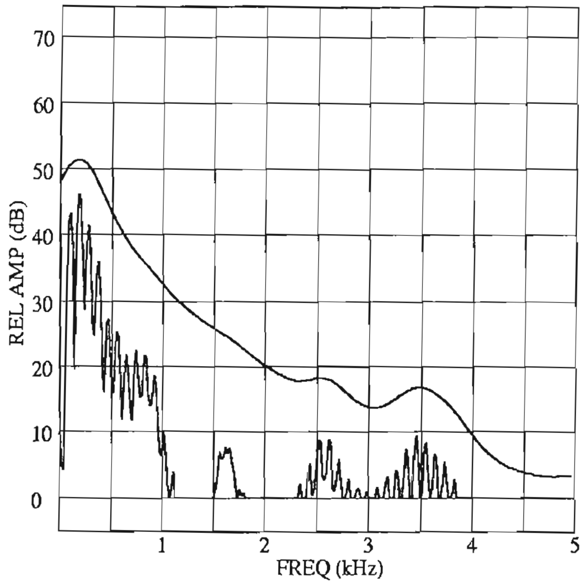

Figure 2-22: DFT Spectrum (Hamming window 39.0 ms long) within the nasal coda at 300 ms of the synthesized tán with one pole-zero pair

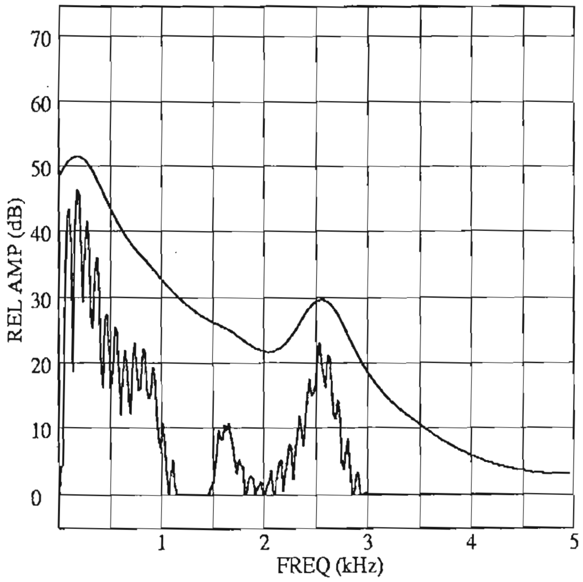

Figure 2-23: DFT Spectrum (Hamming window 39.0 ms long) within the nasal coda at 300 ms of the synthesized tán with two pole-zero pairs

2.5 Perception Test

The perception test was designed for two purposes. First, it was an identification test, where the listener was asked to identify whether the coda to the nasal syllable was [n] or [ŋ]. Second, it provided a measure of the perceptual changes caused by manipulating the number of pole-zero pairs; listeners were asked to rate how natural the ending of the nasal syllable sounded to them on a scale of 1 to 5.

2.5.1 Test Creation

All three iterations of each V-N syllable were synthesized with zero, one, and two pole-zero pairs, totaling 54 synthesized utterances. However, because syntheses of different iterations of the same nasal syllable varied slightly from one another, the best two syntheses of each word, based on the author's judgments, were chosen to be stimuli in the perception test. Therefore there were a total of 36 synthesized utterances in the perception test. In order to provide a basis of comparison for the synthesized syllables, the 12 original utterances were also included in the perception test, for a total of 48 distinct stimuli.

These stimuli were then inputs to a program called MAKETAPE, which outputs an audible string of randomized stimuli based on the user's specifications. This audible output may then be recorded on cassette tape. MAKETAPE generates a file which includes the order in which stimuli are presented. This information was used later in order to sort subjects' responses.

For the creation of this perception test, the 48 stimuli were the inputs. There were a number of factors were considered when creating this perception test. The following were the specifications of the perception test in this study. First, each stimulus was presented to the test subject twice in succession, with 1.5 seconds of silence between them. Second, the test subject required time in which to think about and record his or her responses. In this test, the subject had two tasks; he or she must have identified the nasal coda and also rated the coda. To accommodate this, a silence of 3.5 seconds was inserted after each stimulus. Third, in order to gauge the consistency of the test subject's responses, each stimulus was presented four times, thus resulting in four blocks of 48 stimuli each. Fourth, test subjects require a few stimuli before he or she becomes acclimated to the process of taking the perception test. The first eight stimuli were allocated for this purpose, and were repeated seamlessly at the end of the test. Lastly, in order to help the test subject keep his or her place during the test, a slightly longer pause of 6 seconds was placed after every tenth stimuli.

Two perception tests were produced in the manner described above. Each test had four blocks of 48 stimuli each, plus the 8 duplicate stimuli, totalling 200 stimuli. In the first test, the utterance was simply the nasal syllable in isolation. The duration of this test was approximately 20 minutes. The second test, however, consisted of utterances in which were the nasal syllable was reinserted into its carrier phrase, "shūo CVN bȧ." The duration of this test was approximately 29 minutes.

2.5.2 Conducting the Test

The tests were administered to four native speakers of Mandarin Chinese. All four test subjects were native to Beijing, China. The subjects took the tests individually, starting with the first perception test of the nasal syllable in isolation and then, following a short break, ending with the second perception test of the nasal syllable reinserted into its carrier phrase. The following test-taking procedure was followed for each test subject. The subject was first presented with an instruction sheet, included in Appendix C.2, and was given time to read it over carefully. The instruction sheet presented the six possible test words in traditional Chinese characters as well as in their English phonetic spellings which have been used throughout this document. The format of the test as well as the tasks of the listeners were explained on the instruction sheet. It was explained that they should use the provided answer sheet, an excerpt of which is included in Appendix C.1, to identify whether the ending of the test word was [n] or [ŋ] and to rate the quality of the ending on a scale of one to five, shown in Table 2.2. The subject needed only to circle [n] or [ŋ] as well as a ranking number: 1, 2, 3, 4 or 5. Once the subject indicated that he or she had understood the instructions, the first listening test began.

Table 2.2: Quality rating scale

highly unnatural	unnatural	moderately unnatural	slightly unnatural	natural
1	2	3	4	5

The test tape was played with a Yamaha K-1000 casette deck and was heard by the listener through a set of headphones. The first listening test was conducted without interruption for the duration of the 200 utterances. Once the first test was complete, the subject was given the opportunity to take a short break, after which testing was resumed. The second test was administered in the same manner as the first one, except that the instruction sheet, included in Appendix C.3, now explained that the test word would be inserted into the carrier phrase "shūo _____ bȧ."

Chapter 3

Perception Test Results

The data collected from the perception test were analyzed and plotted in the form of line graphs. The results were plotted to show the relation between the quality of the nasal coda and the number of pole-zero pairs. They convey information about the velar vs. alveolar distinction and the effects of the carrier phrase in the two perception tests. These data, however, should be taken with the understanding that there are several possible sources of error in this study.

3.1 Perception Test Analysis and Results

The perception test responses for each of the four subjects were sorted and analyzed. Even with only four subjects, the data were extensive, with a total of 1600 data points. The data were analyzed and organized to illustrate results based on the number of pole-zero pairs, the alveolar vs. velar distinction, and the effect of the carrier phrase.

3.1.1 Data Analysis

The perception test responses for each test subject were analyzed for misidentifications and ratings. The responses to the first eight stimuli, which were repeated at the end of each perception test, were discarded and not included in the data analysis. The data were then sorted by utterance type, and analysis of the ratings of the

Table 3.1: Number and average rating for utterances in each identification category for utterances in isolation

	perceived [n]		perceived [ŋ]	
	#	rating	#	rating
intended [n]	327	3.9	3	4.0
intended [ŋ]	57	2.6	381	3.5

Table 3.2: Number and average rating for utterances in each identification category for utterances within the carrier phrase

	perceived [n]		perceived [ŋ]	
	#	rating	#	rating
intended [n]	349	4.0	3	4.3
intended [ŋ]	35	3.2	381	3.8

misidentified utterances was conducted. As shown in Tables 3.1 and 3.2, there are four different possible identification categories. The listener may identify the nasal coda as an [n] when it was intended as an [n] or he or she may identify the coda as [n] when it was intended as an [ŋ]. Similarly, there are two possibilities when the listener identifies the utterance as having an [ŋ] coda. These tables show that when an intended [n] coda is perceived as an [ŋ], the average rating over all stimulus types and all four listeners is higher than the ratings for correctly identified utterances. When an intended [ŋ] coda is perceived as an [n], however, the average rating is lower than those for the correctly identified utterances. Because of the significant number of misidentified codas, the data must be adjusted to account for the varying ratings.

To eliminate the effects of the ratings of misidentified nasal codas, the ratings for these responses were changed to a "1" rating of "highly unnatural" and then averaged in to the other ratings. These adjusted ratings were the ones used in all of the subsequently presented data analysis. An average value of the nasal coda rating, which will be called the *total rating*, was calculated over all listeners and iterations of an utterance. The total rating was then used as the basis of comparison among utterances of the same syllable.

3.1.2 Results

Pole-zero pairs

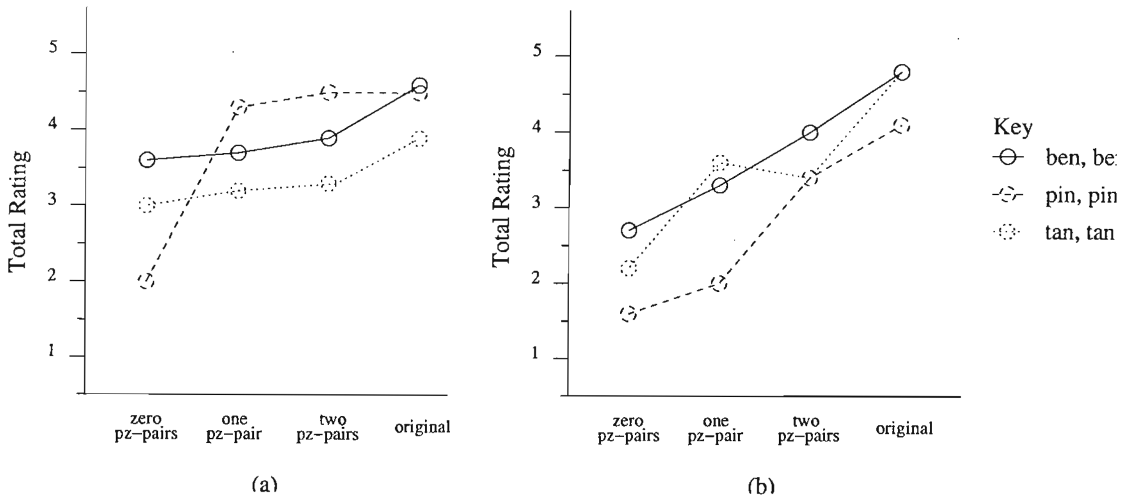

Figure 3-1: Average rating for each utterance for the first perception test with the stimuli presented in isolation. (a) alveolar nasals, (b) velar nasals

The plots of the total rating for each utterance are shown in Figure 3-1 for the first perception test and in Figure 3-2 for the second perception test. Plot (a) in each figure shows the total ratings for the alveolar nasal utterances bēn, pín, and tán while plot (b) shows the data for the velar nasals bēng, píng, and táng. The raw data for these plots are provided in Appendix A, Tables A.1 and A.2.

From these plots, it can be seen that in all but two cases (one from each perception test), there was successive improvement in the perception of the quality of the nasal syllable from zero to one pole-zero pair and from one to two pole-zero pairs. In all cases in both tests, either one or two (or both) pole-zero pairs were better than none.

Table 3.3 shows the number and percentage of misidentified nasal syllables for each different type of stimulus. The data show that for both perception tests, the syntheses with no pole-zero pairs were most often misidentified with the incorrect nasal coda. The 48 repetitions of stimuli with no pole-zero pairs thus were misidentified 70.8% of the time in isolation and 58.3% of the time in the carrier phrase. The clear trend in the data is a marked decrease in misidentifications with each additional pole-zero pair.

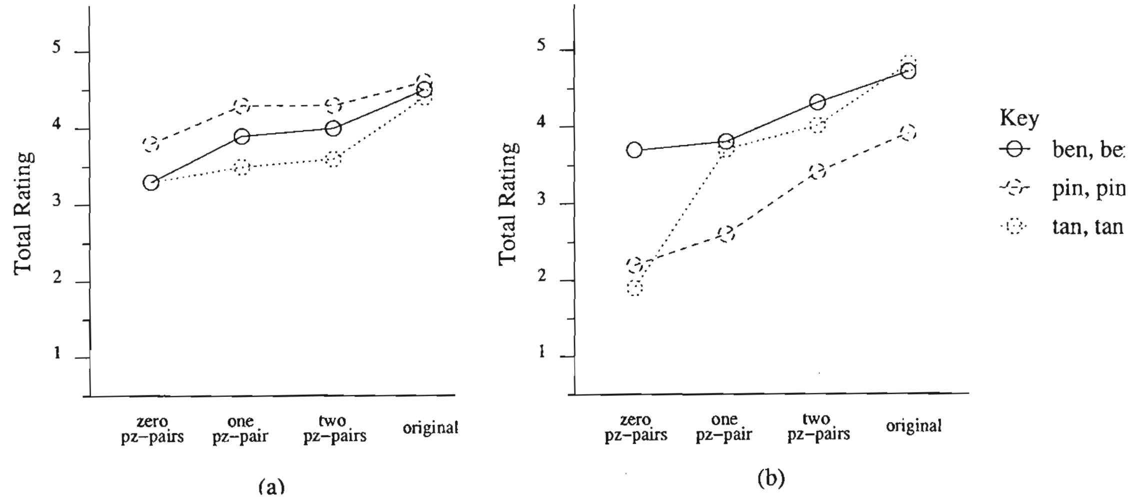

Figure 3-2: Average rating for each utterance for the second perception test with the stimuli embedded in the carrier phrase. (a) alveolar nasals, (b) velar nasals

Table 3.3: Identification data: number and percentage of misidentified utterances for each type of stimulus

	Original		0 pz-pairs		1 pz-pair		2 pz-pairs	
	#	%	#	%	#	%	#	%
Test 1	4	8.3	34	70.8	16	33.3	5	10.4
Test 2	0	0	28	58.3	8	16.7	2	4.2

This trend is more easily seen in the graphical representation shown in Figure 3-3. Here, it is clear that the decrease in misidentified nasal codas is significant with the addition of one and two pole-zero pairs, but that the identification errors associated with the syntheses with two pole-zero pairs were only slightly more prevalent than those of the original utterances.

These data show that each successive pole-zero pair has a positive effect on both the perceptual quality and incidence of correct identification of nasal codas. Each additional pole-zero pair seems to increase perceptual quality and decrease misidentification of the synthesized nasals. Moreover, the syntheses with two pole-zero pairs were, in many cases, comparable to the original utterances in both perceptual quality and percentage of correcty identified utterances.

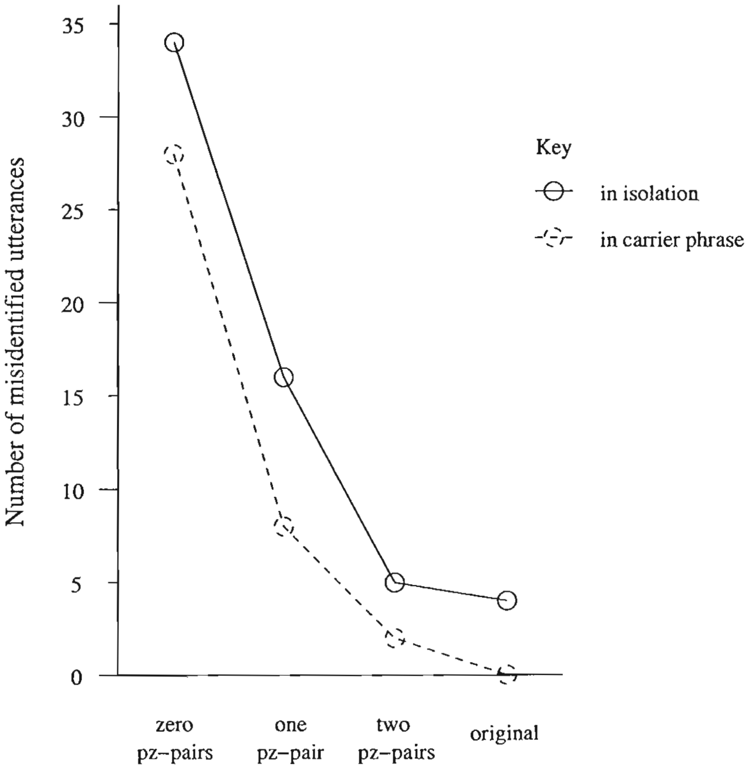

Figure 3-3: Number of identification errors for each type of stimulus

Table 3.4: Increase in total rating for each additional pole-zero pair

	Test 1		Test 2	
	Alveolar	Velar	Alveolar	Velar
zero to one pz pair	0.8	0.8	0.5	0.8
one to two pz pairs	0.2	0.6	0.1	0.5

Alveolar vs. Velar

In Figures 3-1 and 3-2, a difference in the individual effects of the first and second pole-zero pair can be seen. Qualitatively from these graphs, there is more of an increase in average rating with the addition of the second pole-zero pair for velar than for alveolar syllables. Equivalently, the alveolar nasals do not seem to have as much perceptual improvement with the addition of the second pole-zero pair as do the velar nasals. In order to quantify this observation, the increase in rating for each additional pole-zero pair was calculated. A decrease in rating was simply averaged into the total with the other ratings. These data are tabulated in Table 3.4.

The data show that both nasal codas in isolation and within the carrier phrase experienced a greater increase in total rating with the introduction of the low-frequency pole-zero pair than with the introduction of the high-frequency pole-zero pair. The low-frequency pole-zero pair thus appears to be very significant in the perceptual quality of these nasal codas. Additionally, with the incorporation of the second (high-frequency) pole-zero pair, the velar nasal codas in both perception tests increased more in average rating than did the alveolar nasals. Thus with respect to the velar-alveolar distinction, the data show that the high-frequency pole-zero pair could be perceptually more important for the velar nasal than for the alveolar nasal.

Isolation vs. Carrier phrase

There are also data regarding the effect of the carrier phrase in the perception of these Mandarin Chinese nasal syllables. For each listener, an *average rating* was calculated. The average rating was simply the mathematical average of all of the rating scores

Table 3.5: Average rating data for each test subject

Subject	1	2	3	4
Test 1	3.6	3.4	3.5	4.0
Test 2	4.0	3.6	3.6	4.0

Table 3.6: Identification data: number of misidentified stimuli for each test subject

Subject	1	2	3	4	Average
Test 1	20	13	8	19	15
Test 2	12	12	1	13	9.5

which a single listener provided. This average rating would provide a measure of how high or low, on average, an individual test-taker rated the utterances. Table 3.5 shows that the average rating increased in the second perception test for three out of the four listeners. This slight improvement of perceptual quality is expected when the utterance is reinserted into the carrier phrase "shūo CVN bȧ" because of the effects of coarticulation. In this study, coarticulation of the nasal coda with the carrier phrase would have resulted from the physiological effects of the labial closure of the /b/ in "bȧ" with the preceding velar or alveolar nasal.

The effect of the carrier phrase can also be examined by looking at the number of misidentified utterances in each perception test. The average number of misidentified utterances decreased by 5.5 from 15 in the first perception test to 9.5 in the perception test which included the carrier phrase. Table 3.6 summarizes this data. The carrier phrase increases the correct identification of utterances by 2.8%, from 92.2% to 95.0%. The positive effect of the carrier phrase on the identification of the nasal codas can also be seen in Figure 3-3. The utterances which were presented in isolation were consistently misidentified more often than those which were presented in the carrier phrase. As with the improvment in average rating for the utterances in the carrier phrase, this improvement in identification is an expected byproduct of coarticulation. The improvements in identification and average rating suggest slight coarticulation of the nasal coda with the following syllable "bȧ."

3.2 Sources of Error

There are several sources of error in this study. Most could have been remedied if there were enough time and resources. The sources of error include personal bias, differences in test subjects, and the need for more data.

3.2.1 Bias

The perception tests were developed to test the quality of the nasal codas in standard Mandarin Chinese. As will be seen in the Section 3.3, dialectal differences of both the speaker and the listener can greatly affect the study. As the person conducting this study, my own dialectal variations could have played a role in the manner in which I synthesized the utterances. My own dialect of Mandarin is a mix of Taiwanese and American accent. An example of a clear difference in the pronunciation of CVN syllables is the syllable bēng, which Taiwanese speakers of Mandarin would pronounce as [bʌŋ]. This is strikingly dissimilar to the standard [bɛŋ] pronunciation. Additionally, because I grew up in the United States, my own perception of these nasals could have affected the synthesis in a negative way. That is, it is possible that my judgment is not as sensitive to the important cues for nasals as that of a native speaker of standard Mandarin.

3.2.2 Differences in test subjects

In addition to my personal contribution to the error, the test subjects who took the two perception tests also are a source of error. Although they all speak the same dialect and presumably have similar judgement regarding nasal codas in Mandarin Chinese, the way in which they rate the nasal codas varies quite a bit. For instance, the average utterance rating in the first perception test for one listener was 4.0 whereas it was 3.4 for another listener. The complete set of data is shown in Table 3.5. This discrepancy in rating comes about because one listener may be more or less strict about the quality of the coda than another listener. Although it was not done here, this type of error can be slightly corrected by renormalizing quality ratings by the average rating of each listener.

It can also be seen from the raw data of each listener that some test subjects are more consistent than others. Some listeners rate the exact same utterance with four different ratings while others are able to consistently rate utterances with one or two different ratings. This type of error might not be easily corrected, but could be remedied with increased repetition of each unique utterance within the test.

3.2.3 Need for more data

Lastly, in order to arrive at more substantiated conclusions, the perception test would need to be conducted on more listeners. Four listeners, although enough to give a rough idea of trends, are not substantial enough to make broad claims. Additionally, this study only examined the utterances made by one male speaker. In order to create a more comprehensive study, the utterances of multiple male and female speakers should be recorded and synthesized.

3.3 Additional Observations

In the course of designing this study of nasal codas, many intermediate syntheses and trials were explored. We describe here some observations regarding these different courses of study which were performed. These observations point to the need for further research.

3.3.1 Variations in Speakers

The syntheses performed in this study were copy-synthesized from the utterances of a male speaker. With this particular speaker, synthesis of his voice quality was relatively simple because his voice was steady and clear. In preparation for this study, the voices of two female speakers were also recorded for synthesis. The voice of the first female seaker, in addition to being of higher pitch than the male speaker used in this study, was slightly uneven and breathy, proving to be a more difficult voice quality

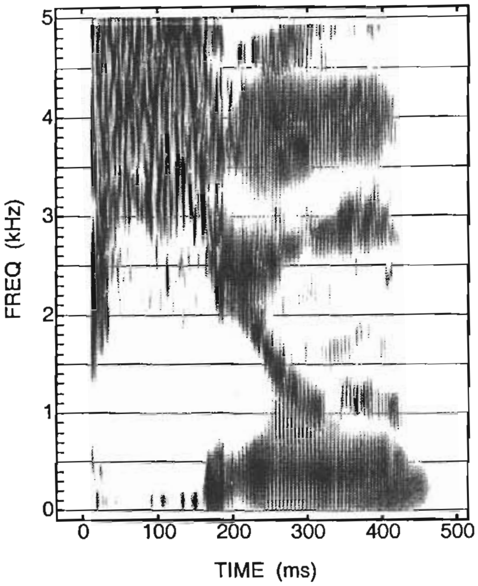

Figure 3-4: Spectrogram of píng for a speaker from southern China

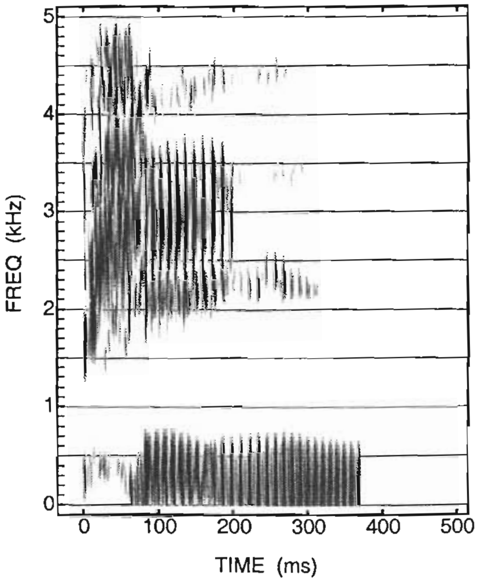

Figure 3-5: Spectrogram of píng for a speaker of standard Mandarin Chinese

to synthesize. With such a voice, this study would have been much more difficult, as the voice quality of an utterance is an important cue to the quality synthesis.

Some nasal utterances of the second female speaker were also recorded and synthesized. This female speaker was a native speaker of Mandarin, but spoke with the native accent of her hometown in southern China. It is widely known that the various accents which exist in southern China can differ greatly from the standard Mandarin Chinese which is spoken in northern China, namely in Beijing. For this particular speaker, the production of the syllable píng varied greatly from the production of the syllable by our original male speaker. The major difference between the two utterances can be seen clearly in their spectrograms, shown in Figures 3-4 and 3-5. The female speaker created a diphthong of /iʌ/ instead of the vowel /i/ while the male speaker produced a more consistent /i/ vowel.

3.3.2 Variations in Listeners

In addition to the effects of dialect on spoken nasal utterances, there were also effects of dialect on the perception of nasal utterances. The perception tests were also administered to two individuals who were native speakers of dialects from southern China.

Brief analysis of their data indicated that they had significant trouble identifying whether the nasal was [n] or [ŋ]. In most cases, these two subjects incorrectly identified all or most of the iterations of any given utterance. Moreover, even in these misidentified cases, the subjects still recorded mostly the relatively high quality ratings of 4 or 5. When the correct identifications were made, there was no clear relationship between average rating and number of pole-zero pairs.

Chapter 4

Conclusion

4.1 Discussion and Conclusions

The data presented in Chapter three allow for three main conclusions regarding the effects of additional pole-zero pairs and carrier phrases on the perception of nasal codas in Mandarin Chinese.

First, it appears that in this study, the carrier phrase had a small positive effect on the perception of the synthesized syllable. On average, there was approximately 2.8% of an increase in correct identification corresponding to an increase of 5.25 more correctly identified stimuli when they were presented within the carrier phrase than when they were in isolation. This suggested that there could have been slight coarticulation of the nasal with the carrier phrase "shūo CVN bȧ." This was not unexpected when the place of articulation at the end of a syllable differed from that of the following phoneme. In this case, the labial stop consonant in "bȧ" was probably slightly coarticulated with the velar or alveolar nasal.

Second, from the data shown in this study, the addition of up to two pole-zero pairs into the transfer function resulted in improved perception of the nasal codas. The incorporation of the high- and low-frequency pole-zero pairs led to a more accurate representation of the V-N boundary and nasal murmur spectra. The spectral differences produced by the pole-zero pairs in the low and high frequency bands resulted in varying perceptual qualities. Specifically, the low-frequency pole-zero pair resulted in a greater increase in perceptual rating than the high-frequency pole-zero pair for both alveolar and velar nasals. Because of the greater sensitivity of the human ear to lower frequency bands, this result was not surprising. Additionally, the use of the high-frequency pole-zero pair resulted in greater perceptual quality increase in velar nasal syllables than the alveolar ones. This suggested that the high-frequency pole-zero pair could be a clue to the velar/alveolar distinction.

Third, from the relative total rating values for each utterance in most cases, the spectral effects of the first and second pole-zero pairs did in fact successively increase the perceptual quality of the nasal coda. Also, each successive pole-zero pair decreased the number of misidentified nasal codas, suggesting that the incorporation of the high-frequency pole-zero pair in synthesis was not only important for velar/alveolar identification, but also for the perceptual quality of the nasal coda. The increase in perceptual quality and incidence of correct identification with either one or two pole-zero pairs suggests that the use of pole-zero pairs may be necessary for improved syntheses of nasal codas in Mandarin.

4.2 Further Study

There are a number of improvements that could be made to this study. First, some of the errors in Section 3.2 could be corrected by increasing the number of subjects taking the perception tests and normalizing the data of the test subjects by their individual rating average. Second, more syllables containing these two nasal codas need to be examined. Only with the study of more nasal syllables in the language can generalizations of the acoustic properties of nasals be solidified. Third, improved syntheses could be made. Although the best syntheses came reasonably close in averate rating to the original, given more time it would be possible to synthesize an utterance which would be identical in perception to the orginal.

This study has also touched upon a number of other interesting research areas in the subject of nasal codas in Mandarin Chinese. Each of these courses of further study would offer additional insight into the acoustic characteristics of nasal codas in Mandarin. This study was conducted only with syllables of two out of the four Mandarin tones. In order to properly characterize the role of pole-zero pairs across all nasal syllables in Mandarin, all of the four tones should be examined. Also, studies should be conducted with the remaining nasal codas in Mandarin which do not occur in the [n]-[ŋ] pairs as well as those nasal codas which are preceded by diphthongs. Additionally, it would be interesting to study the effects of dialect which were briefly explored in Section 3.3.

The results of this study will hopefully be a starting point for future research and increased understanding of the acoustics and synthesis of nasal codas in Mandarin Chinese.

Appendix A

Tables

Table A.1: Total Rating data for Perception Test 1

	Alveolar			Velar		
Utterance	bēn	pín	tán	bēng	píng	táng
0 pz pairs	3.6	2.0	3	2.7	1.6	2.2
1 pz pair	3.7	4.3	3.2	3.3	2.0	3.6
2 pz pairs	3.9	4.5	3.3	4	3.4	3.4
Original	4.6	4.5	3.9	4.8	4.1	4.8

Table A.2: Total Rating data for Perception Test 2

	Alveolar			Velar		
Utterance	bēn	pín	tán	bēng	píng	táng
0 pz pairs	3.1	3.8	3.3	3.7	2.2	1.9
1 pz pair	3.9	4.3	3.5	3.8	2.6	3.7
2 pz pairs	4	4.3	3.6	4.3	3.4	4
Original	4.5	4.6	4.4	4.7	3.9	4.8

Table A.3: KLSYN88 Parameters [10]

UI	update interval
NF	number of formants
SS	source switch
RS	random seed
SB	same noise burst
CP	cascade or parallel toggle
OS	output selector
SQ	speed quotient
OQ	open quotient
TL	spectral tilt
FL	flutter
DI	diplophonia
SR	sampling rate of the digitized speech
DU	duration of speech
F0, F1, F2, F3, F4, F5, F6	fundamental and formant frequencies
B1, B2, B3, B4, B5, B6	bandwidths of formant frequencies
DF1	change in F_1 and B_1, respectively, during open portion of a period
A2F, A3F, A4F, A5F, A6F	amplitude of frication at each formant
B2F, B3F, B4F, B5F, B6F	bandwidth of frication at each formant
GV, GH, GF	overall gain of voicing, aspiration, and frication, respectively
AV, AH, AF	amplitude of voicing, aspiration, and frication, respectively
FNZ, FNP	frequency of low-frequency pole-zero pair
FTZ, FTP	frequency of high-frequency of pole-zero pair.
BNZ, BNP, BTZ, BTP	bandwidths of the two pole-zero pairs.
ANV	amplitude of voicing at low-frequency nasal formant
ATV	amplitude of voicing at high-frequency nasal formant
A1V, A2V, A3V, A4V	amplitude of voicing at each formant

Appendix B

Figures

All of the DFT spectra shown in this Appendix were calculated with a 39.0 ms Hamming window.

B.1 DFT spectra and spectrograms of original and synthesized táng

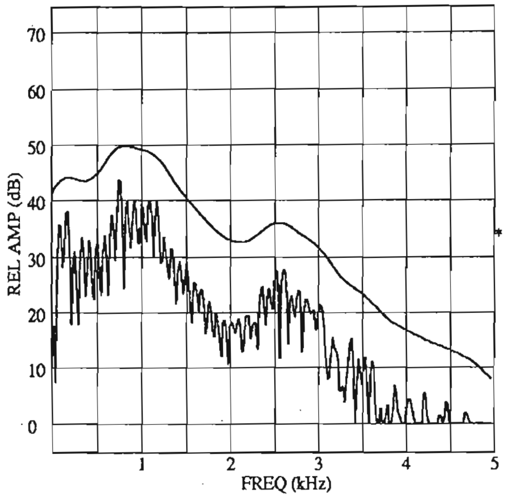

Figure B-1: DFT spectrum of the original táng in the vowel at 175 ms

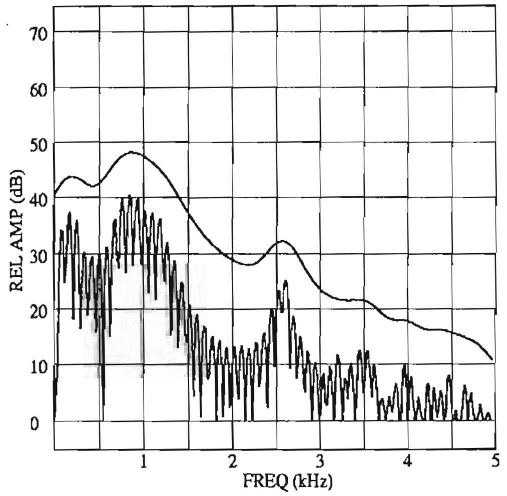

Figure B-2: DFT spectrum of the synthesized táng in the vowel at 175 ms

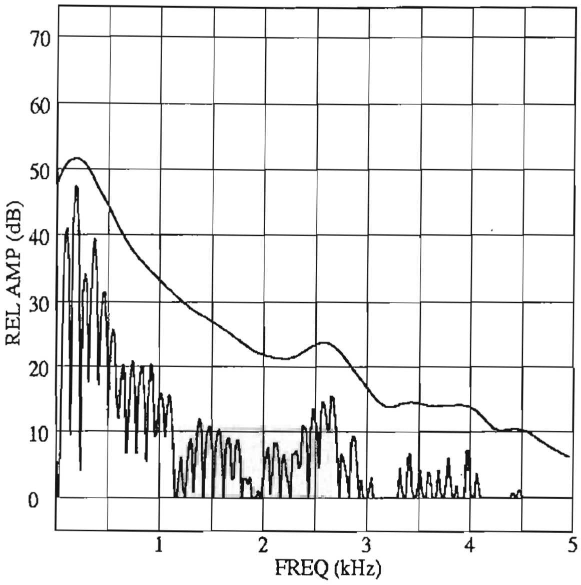

Figure B-3: DFT spectrum of the original táng in the nasal murmur at 300 ms

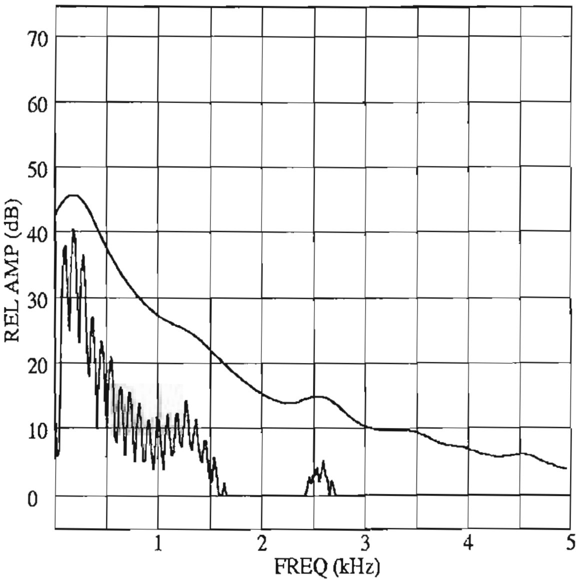

Figure B-4: DFT spectrum of the synthesized táng with no pole-zero pairs in the nasal murmur at 300 ms

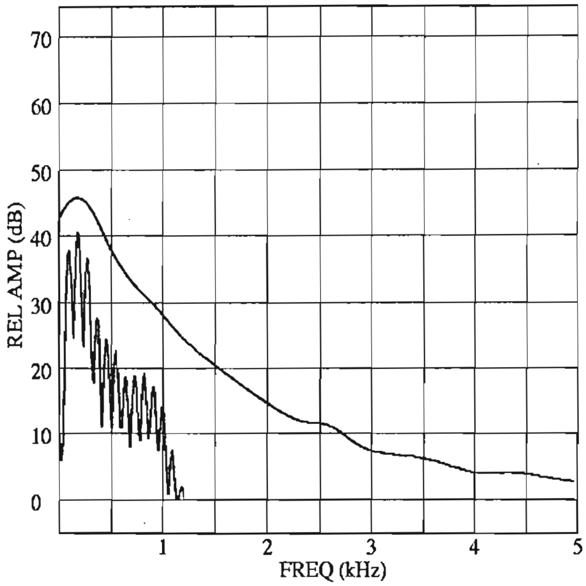

Figure B-5: DFT spectrum of the synthesized táng with one pole-zero pairs in the nasal murmur at 300 ms

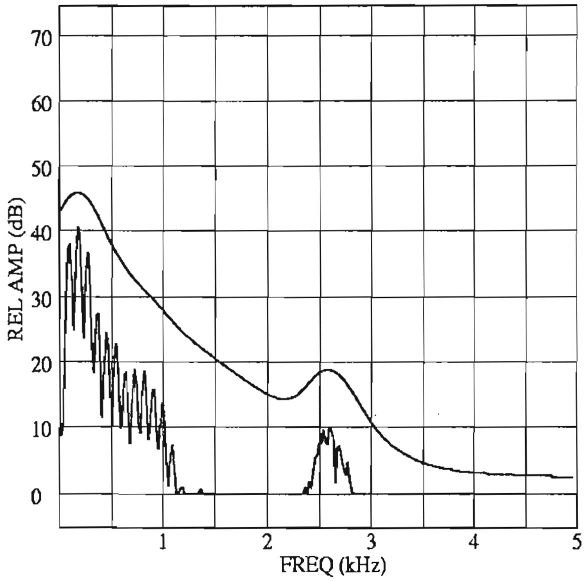

Figure B-6: DFT spectrum of the synthesized táng with two pole-zero pairs in the nasal murmur at 300 ms

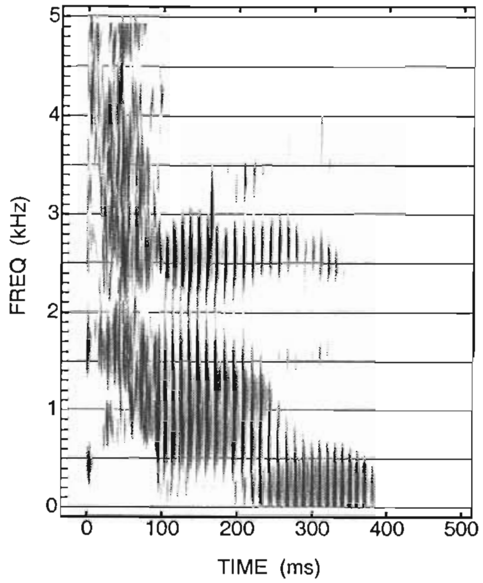

Figure B-7: Spectrogram of original táng

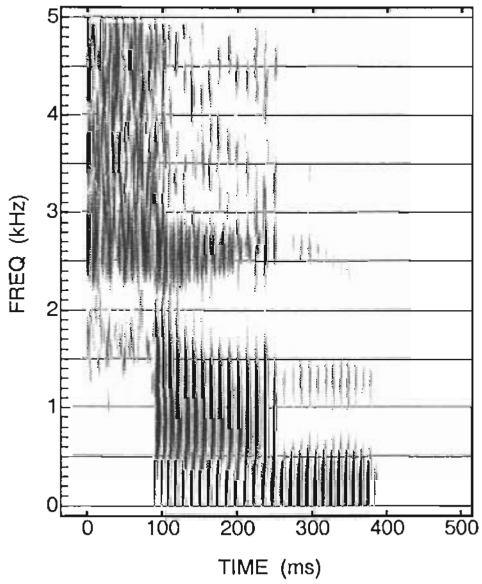

Figure B-8: Spectrogram of synthesized táng with no pole-zero pairs

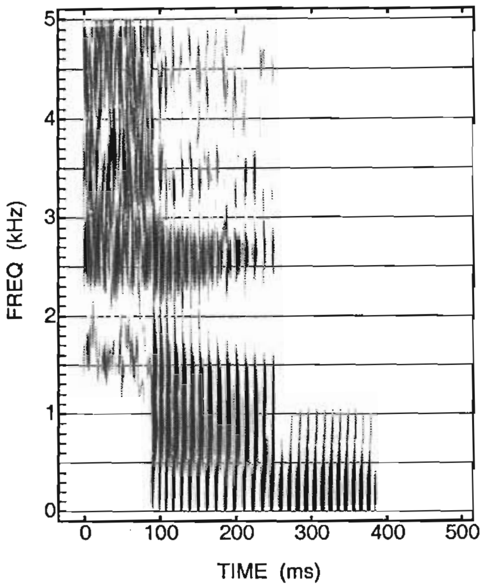

Figure B-9: Spectrogram of synthesized táng with one pole-zero pair

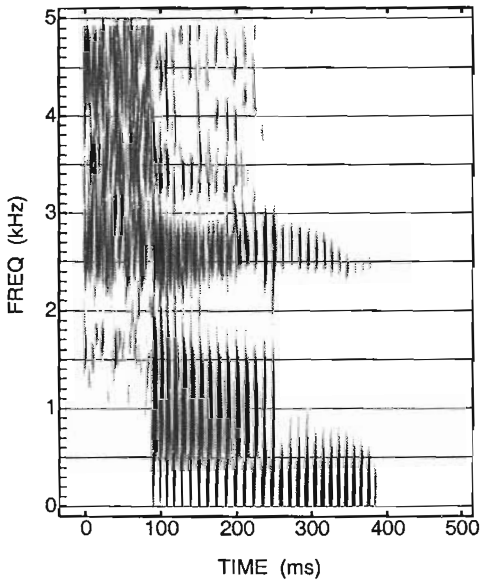

Figure B-10: Spectrogram of synthesized táng with two pole-zero pairs

B.2 DFT spectra and spectrograms of original and synthesized bēn

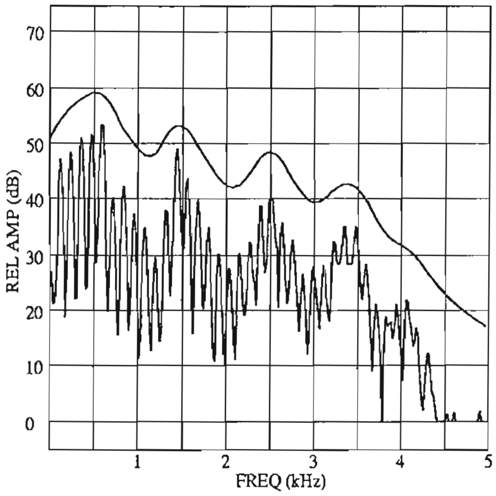

Figure B-11: DFT spectrum of the original bēn in the vowel at 75 ms

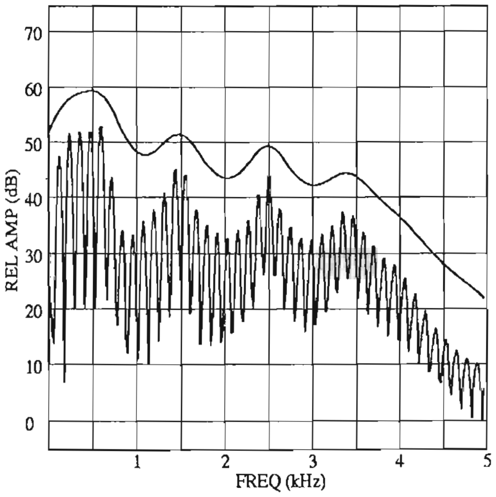

Figure B-12: DFT spectrum of the synthesized bēn in the vowel at 75 ms

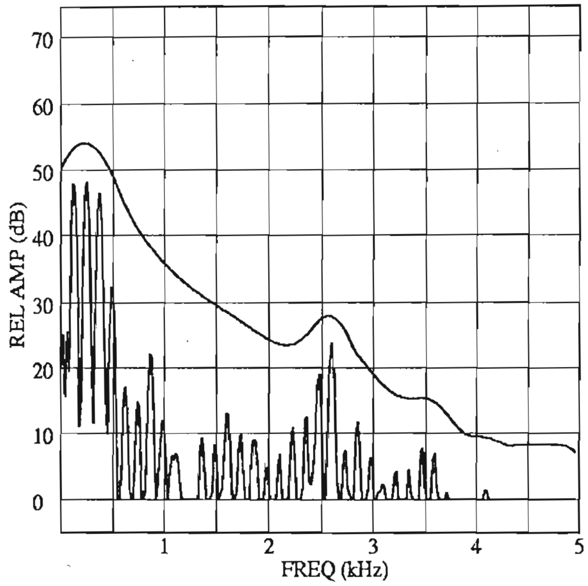

Figure B-13: DFT spectrum of the original bēn in the nasal murmur at 175 ms

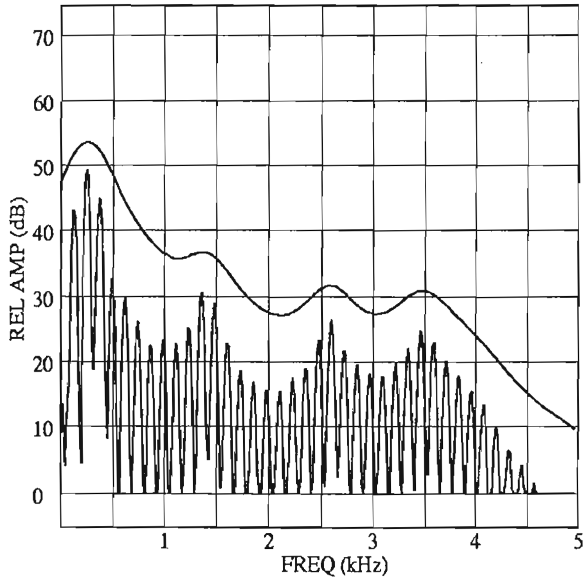

Figure B-14: DFT spectrum of the synthesized bēn with no pole-zero pairs in the nasal murmur at 175 ms

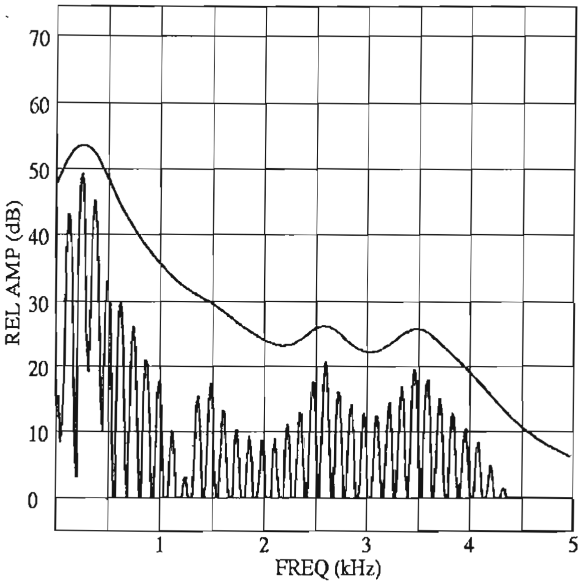

Figure B-15: DFT spectrum of the synthesized bēn with one pole-zero pairs in the nasal murmur at 175 ms

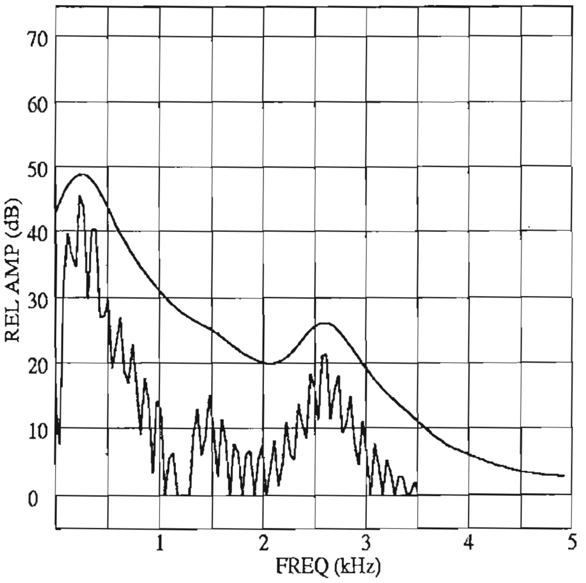

Figure B-16: DFT spectrum of the synthesized bēn with two pole-zero pairs in the nasal murmur at 175 ms

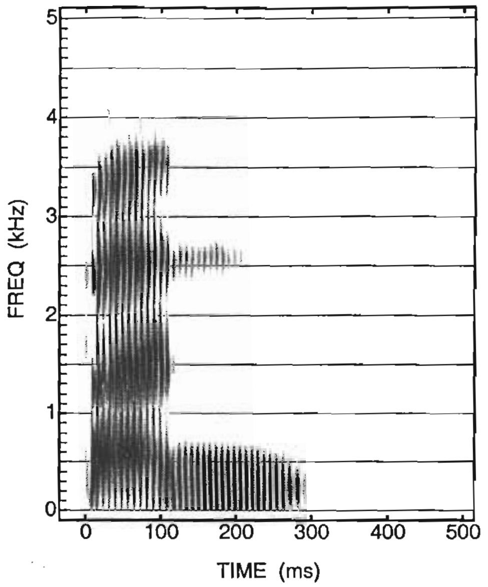

Figure B-17: Spectrogram of original bēn

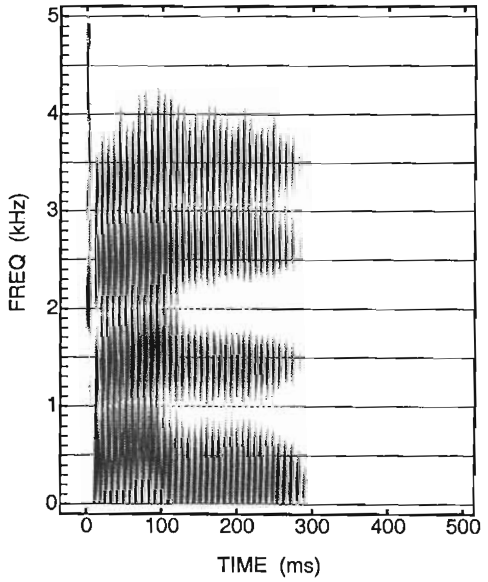

Figure B-18: Spectrogram of synthesized bēn with no pole-zero pairs

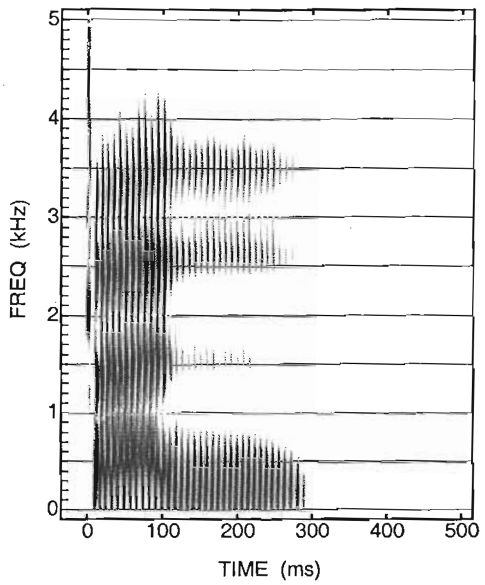

Figure B-19: Spectrogram of synthesized bēn with one pole-zero pair

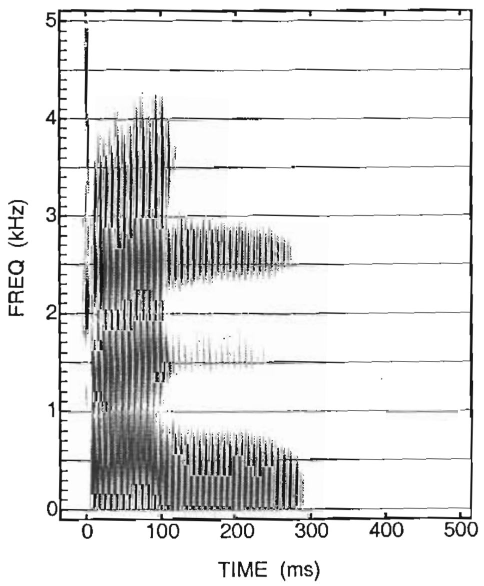

Figure B-20: Spectrogram of synthesized bēn with two pole-zero pairs

B.3 DFT spectra and spectrograms of original and synthesized bēng

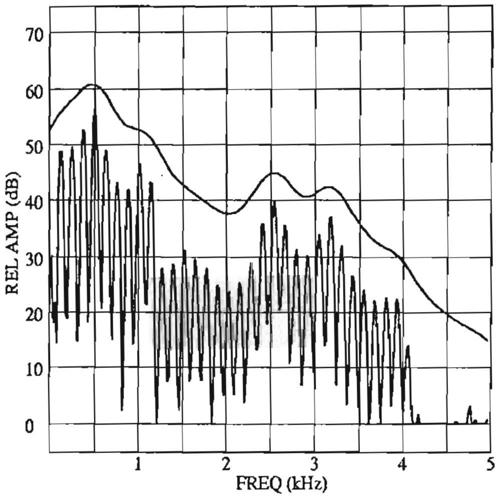

Figure B-21: DFT spectrum of the original bēng in the vowel at 75 ms

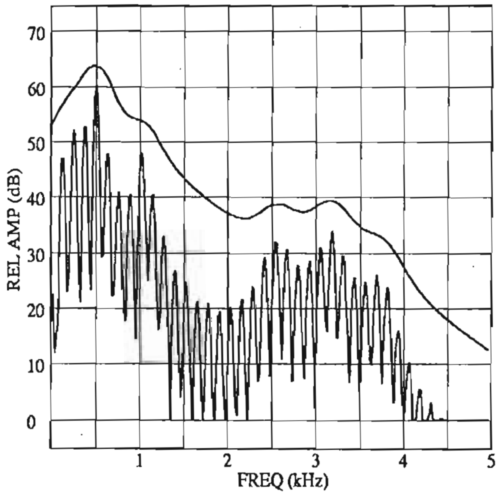

Figure B-22: DFT spectrum of the synthesized bēng in the vowel at 75 ms

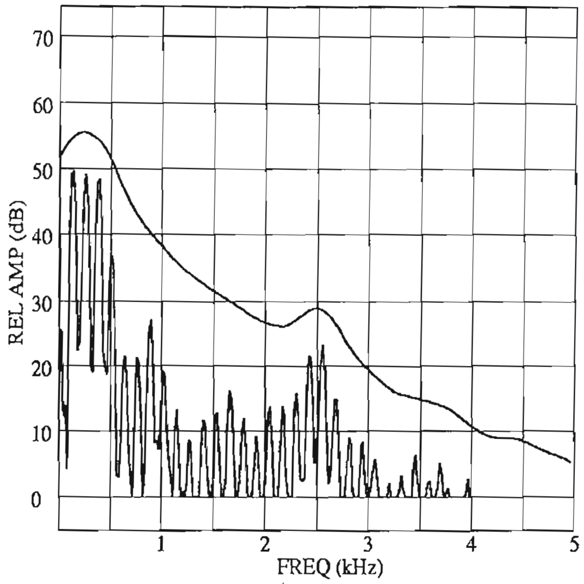

Figure B-23: DFT spectrum of the original bēng in the nasal murmur at 175 ms

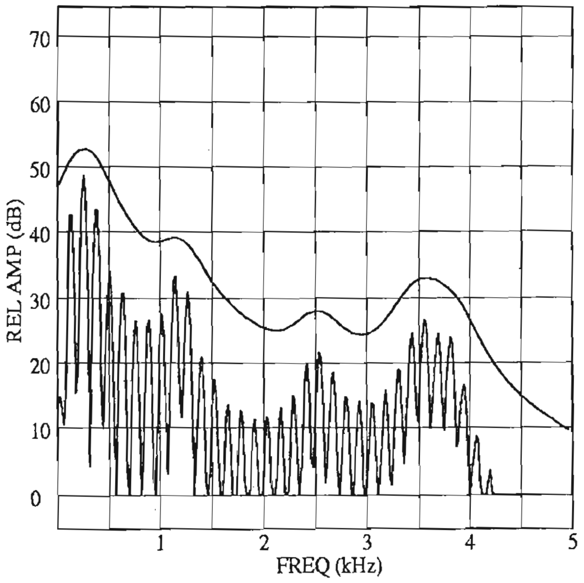

Figure B-24: DFT spectrum of the synthesized bēng with no pole-zero pairs in the nasal murmur at 175 ms

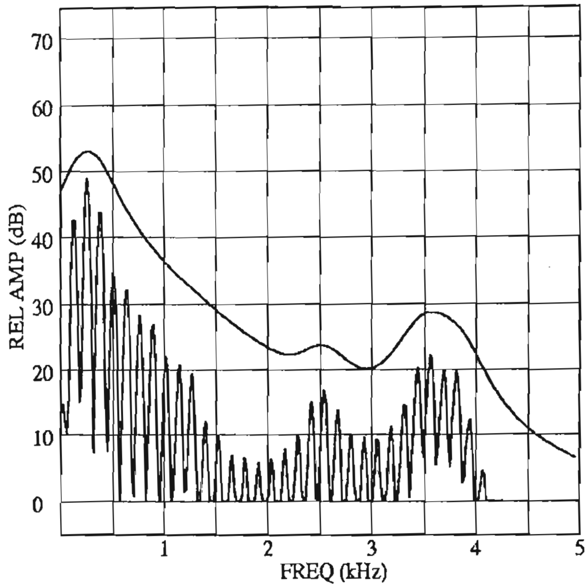

Figure B-25: DFT spectrum of the synthesized bēng with one pole-zero pairs in the nasal murmur at 175 ms

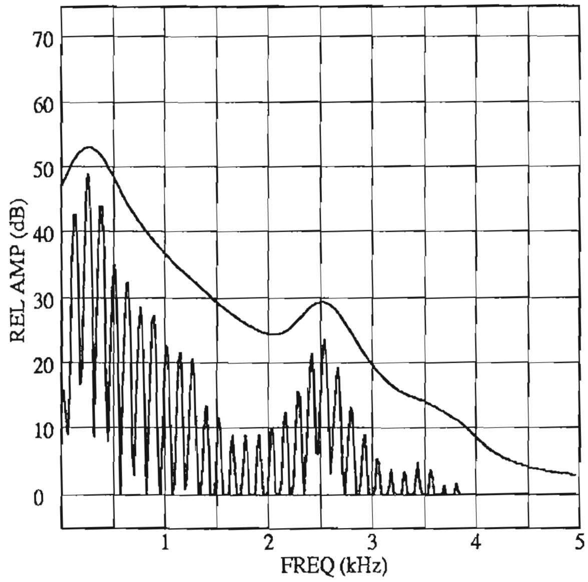

Figure B-26: DFT spectrum of the synthesized bēng with two pole-zero pairs in the nasal murmur at 175 ms

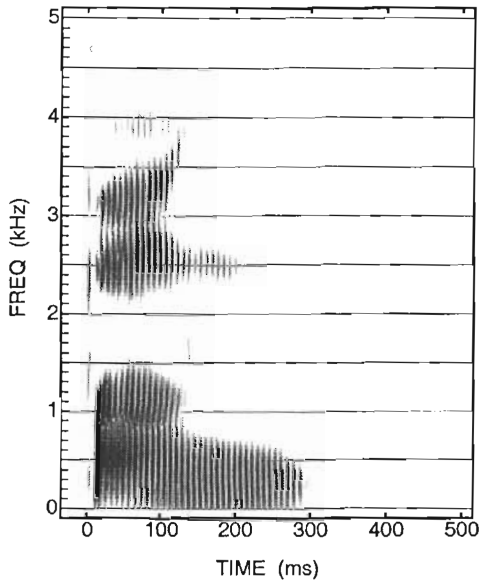

Figure B-27: Spectrogram of original bēng

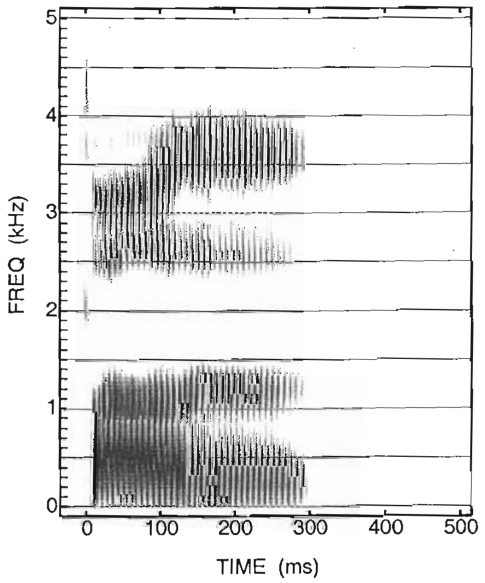

Figure B-28: Spectrogram of synthesized bēng with no pole-zero pairs

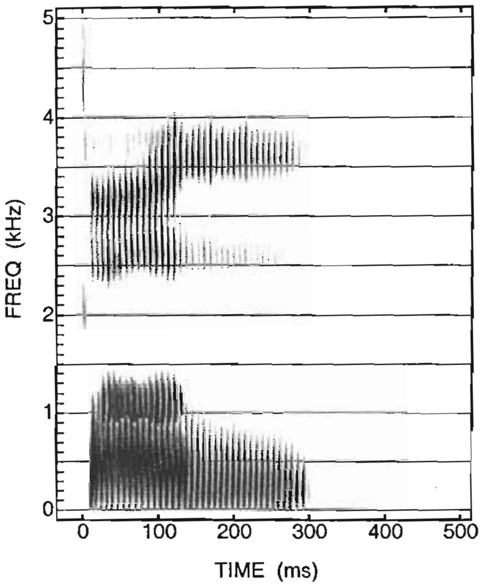

Figure B-29: Spectrogram of synthesized bēng with one pole-zero pair

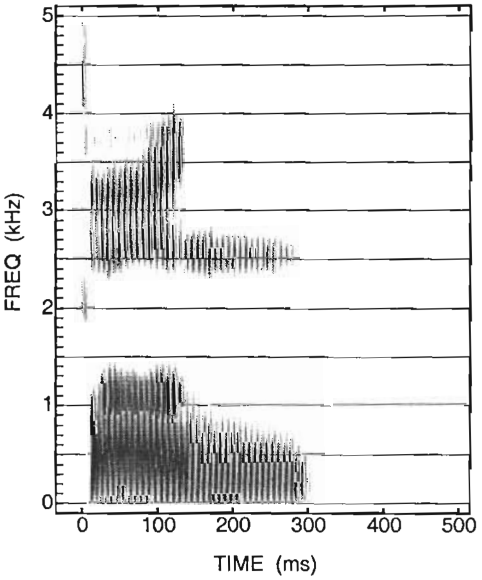

Figure B-30: Spectrogram of synthesized bēng with two pole-zero pairs

B.4 DFT spectra and spectrograms of original and synthesized pín

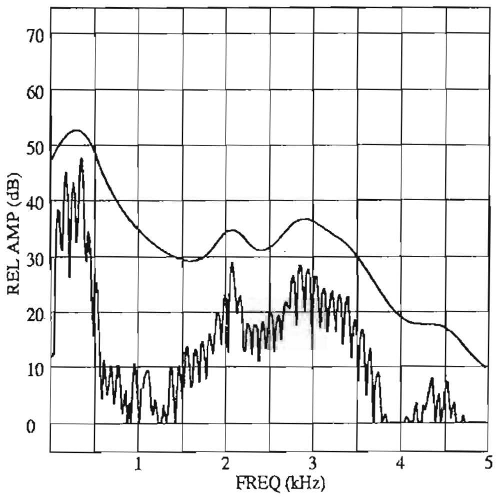

Figure B-31: DFT spectrum of the original pín in the vowel at 75 ms

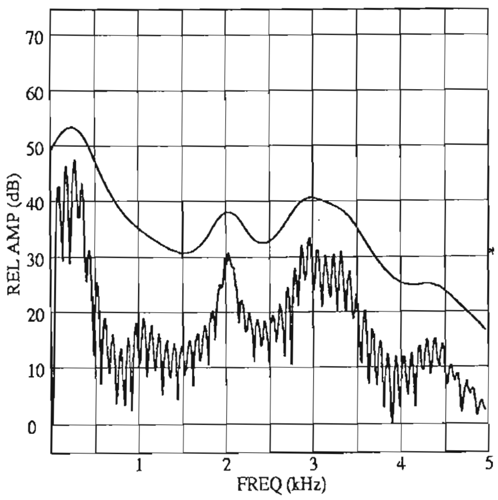

Figure B-32: DFT spectrum of the synthesized pín in the vowel at 75 ms

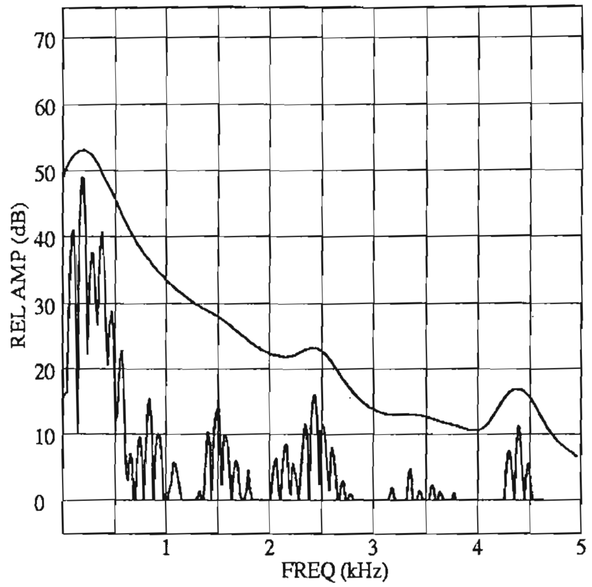

Figure B-33: DFT spectrum of the original pín in the nasal murmur at 175 ms

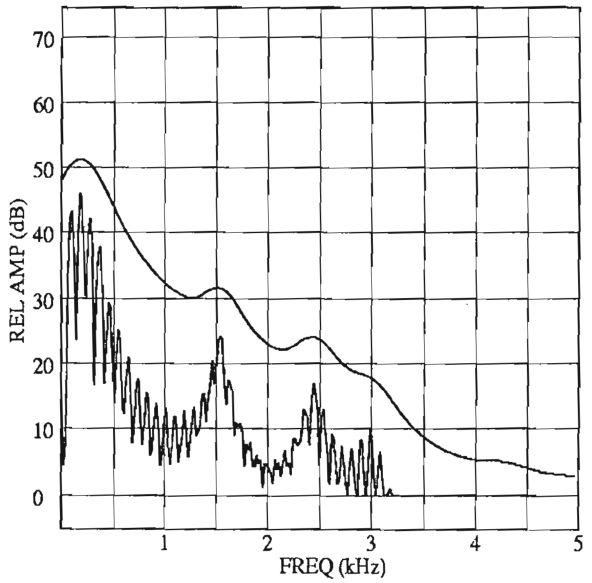

Figure B-34: DFT spectrum of the synthesized pín with no pole-zero pairs in the nasal murmur at 175 ms

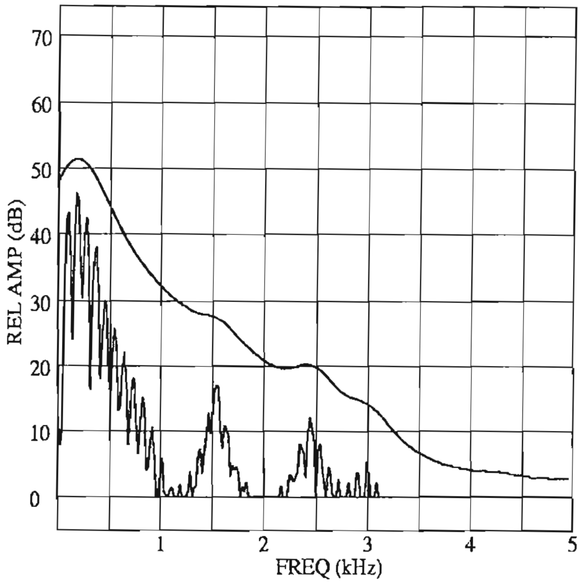

Figure B-35: DFT spectrum of the synthesized pín with one pole-zero pairs in the nasal murmur at 175 ms

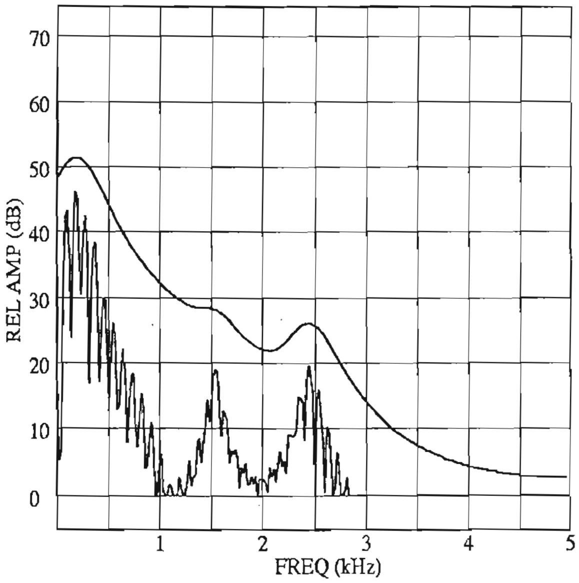

Figure B-36: DFT spectrum of the synthesized pín with two pole-zero pairs in the nasal murmur at 175 ms

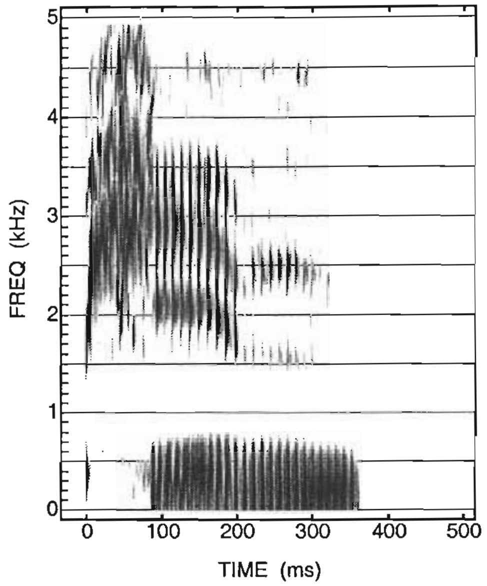

Figure B-37: Spectrogram of original pı́n

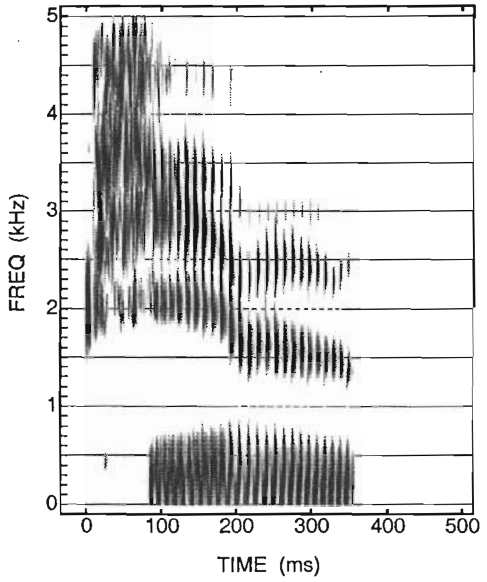

Figure B-38: Spectrogram of synthesized pı́n with no pole-zero pairs

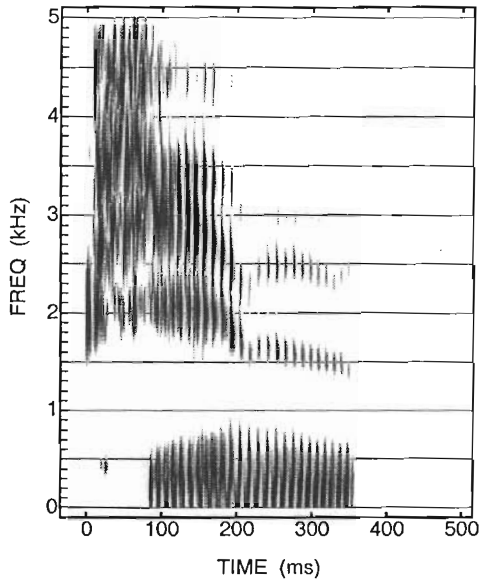

Figure B-39: Spectrogram of synthesized pín with one pole-zero pair

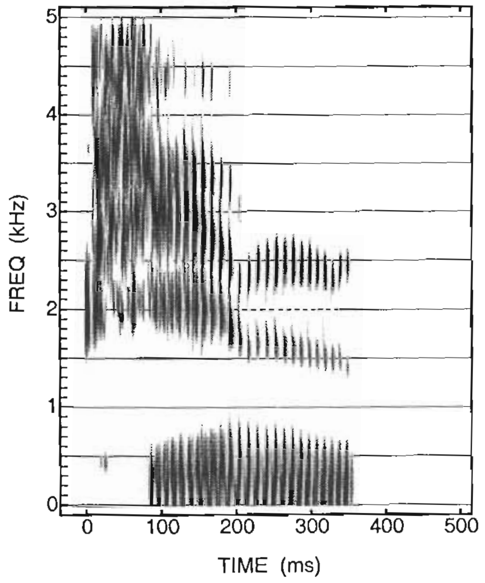

Figure B-40: Spectrogram of synthesized pín with two pole-zero pairs

B.5 DFT spectra and spectrograms of original and synthesized píng

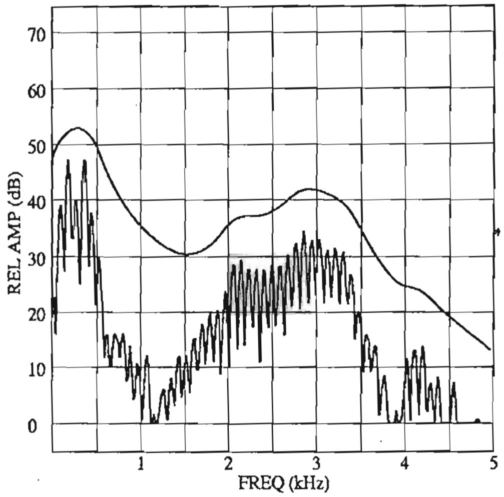

Figure B-41: DFT spectrum of the original píng in the vowel at 75 ms

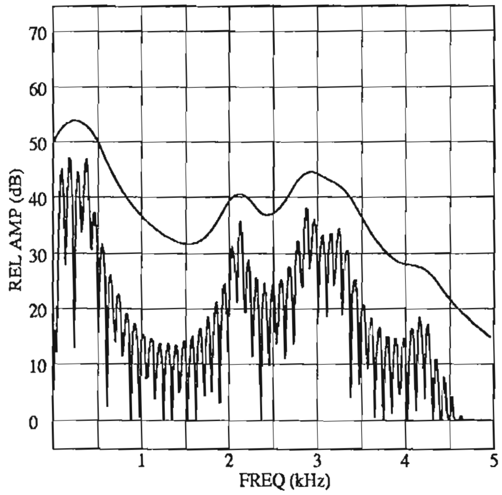

Figure B-42: DFT spectrum of the synthesized píng in the vowel at 75 ms

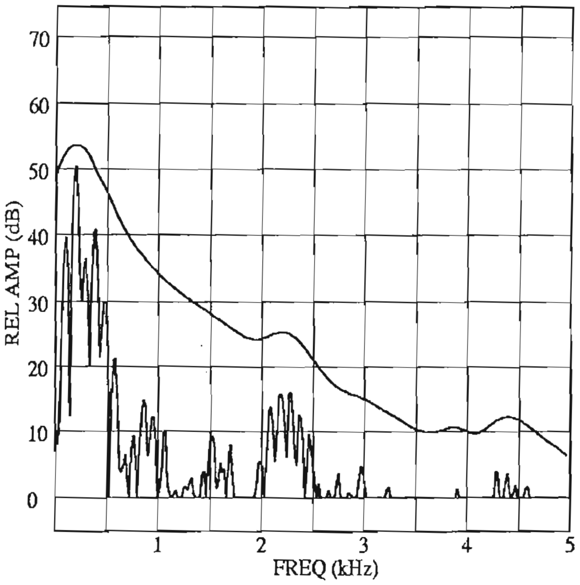

Figure B-43: DFT spectrum of the original píng in the nasal murmur at 175 ms

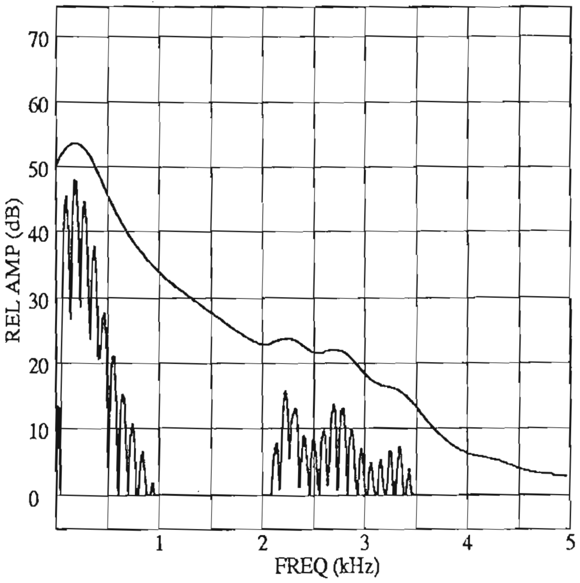

Figure B-44: DFT spectrum of the synthesized píng with no pole-zero pairs in the nasal murmur at 175 ms

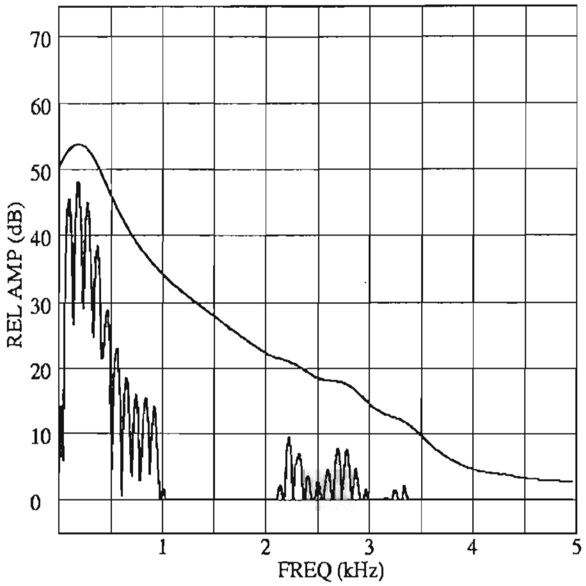

Figure B-45: DFT spectrum of the synthesized pińg with one pole-zero pairs in the nasal murmur at 175 ms

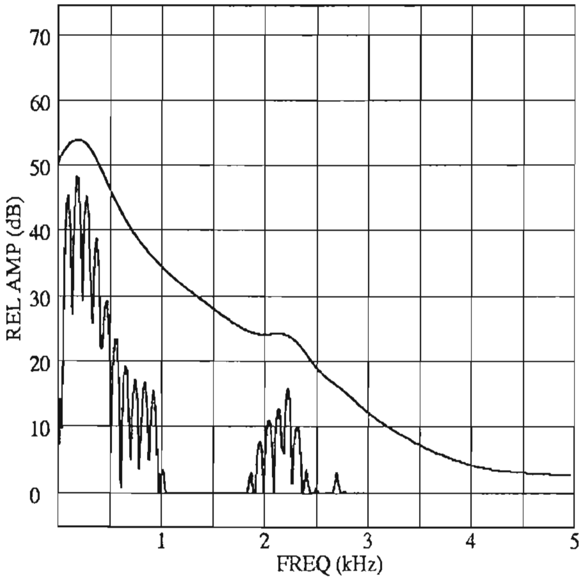

Figure B-46: DFT spectrum of the synthesized pińg with two pole-zero pairs in the nasal murmur at 175 ms

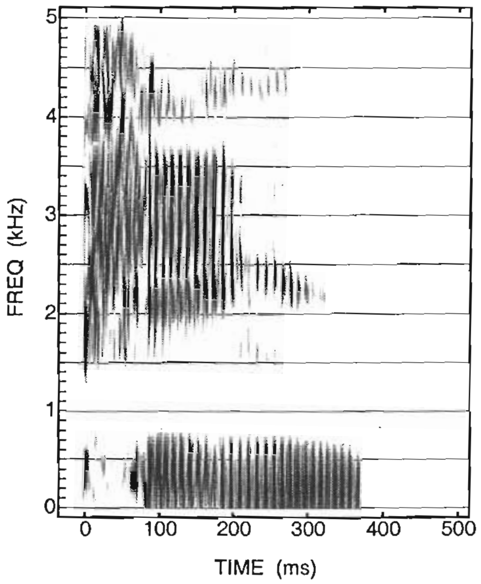

Figure B-47: Spectrogram of original pi&#769;ng

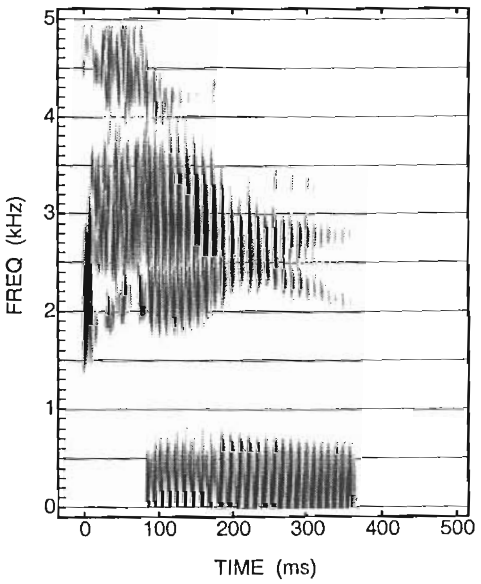

Figure B-48: Spectrogram of synthesized pi&#769;ng with no pole-zero pairs

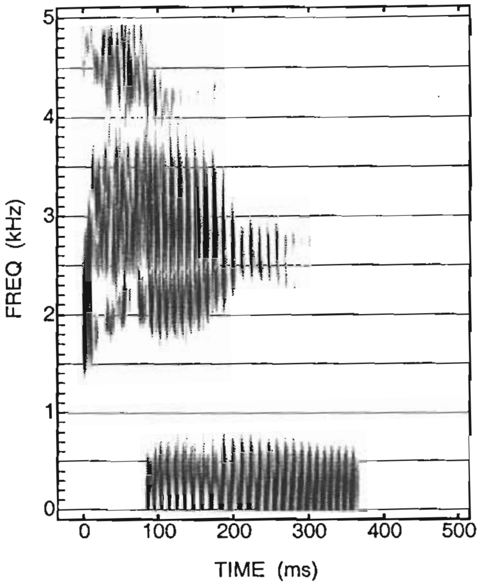

Figure B-49: Spectrogram of synthesized píng with one pole-zero pair

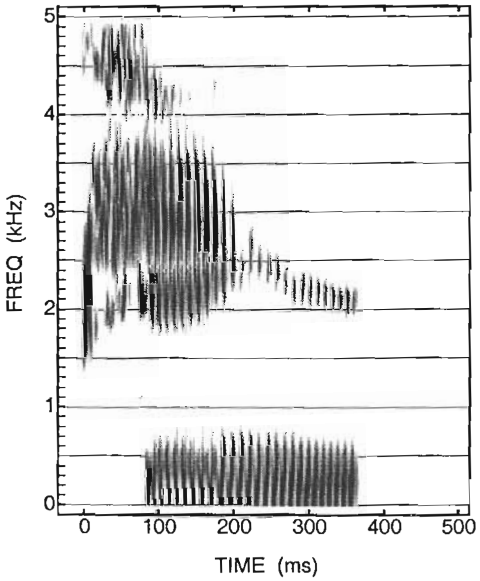

Figure B-50: Spectrogram of synthesized píng with two pole-zero pairs

Appendix C

Perception test documents

C.1 Response Sheet

This is an excerpt from the perception test answer sheet used to record the subject's responses. The original answer sheet had 200 response entries.

	Ending		Quality Rating				
1.	n	ng	1	2	3	4	5
2.	n	ng	1	2	3	4	5
3.	n	ng	1	2	3	4	5
4.	n	ng	1	2	3	4	5
5.	n	ng	1	2	3	4	5
6.	n	ng	1	2	3	4	5
7.	n	ng	1	2	3	4	5
8.	n	ng	1	2	3	4	5
9.	n	ng	1	2	3	4	5
10.	n	ng	1	2	3	4	5

C.2 First listening test instructions

Thank you for participating in this listening experiment. In a few minutes, you will be hearing a series of utterances. Each utterance will consist of one of the following test words:

[Traditional Chinese characters were presented above each pinyin representation]

pín píng bēn bēng tán táng

Your tasks for each utterance are:

1. to identify whether the ending to the test word is [n] or [ng] by circling your response.
2. to rate the quality of the ending of the test word based on the following scale:

highly unnatural	unnatrual	moderately unnatural	slightly unnatural	natural
1	2	3	4	5

You will hear each test word twice in succession.

After a brief pause, the next test word will be played twice.

In order to help you keep your place, there is a slightly longer pause after every tenth utterance.

There are a total of 200 utterances.

C.3 Second listening test instructions

Thank you for participating in this listening experiment. In a few minutes, you will be hearing a series of utterances. The utterances will be of the form

[Traditional Chinese characters were presented above each pinyin representation]

“shūo _____ bȧ”

where the blank will be filled by one of the following test words:

[Traditional Chinese characters were presented above each pinyin representation]

pín píng bēn bēng tán táng

Your tasks for each utterance are:

1. to identify whether the ending to the test word is [n] or [ng] by circling your response.
2. to rate the quality of the ending of the test word based on the following scale:

highly unnatural	unnatrual	moderately unnatural	slightly unnatural	natural
1	2	3	4	5

You will hear each test word twice in succession.

After a brief pause, the next test word will be played twice.

In order to help you keep your place, there is a slightly longer pause after every tenth utterance.

There are a total of 200 utterances.

References

[1] M. Y. Chen. Acoustic correlates of English and French nasalized words. *Journal of the Acoustical Society of America*, 102:2360–2370, 1997.

[2] M. Y. Chen. Acoustic analysis of simple vowels preceding a nasal in Standard Chinese. *Journal of Phonetics*, 128:43–67, 2000.

[3] C. P. Chou, P. Link, and X. Wang. *Oh, China! : Elementary Reader of Modern Chinese for Advanced Beginners.* Princeton University Press, 1997.

[4] Z. Handel. Pinyin review. http://courses.washington.edu/chinese2/pinyin.pdf, December 2001.

[5] H. M. Hanson. Synthesis of female speech using the klatt synthesizer. *Speech Communication Group Working Papers, Research Laboratory for Electronics, MIT*, 10:84–103, December 1995.

[6] J. Harrington. The contribution of the murmur and vowel to the place of articulation distinction in nasal consonants. *Journal of the Acoustical Society of America*, 96:19–32, 1994.

[7] A. S. House. Analog studies of nasal consonants. *Journal of Speech and Hearing Disorders*, 102:2360–2370, 1997.

[8] J. M. Howie. *Acoustical Studies of Mandarin Vowels and Tones.* Cambridge University Press, 1976.

[9] W. Hu. Chinese pronunciation guide. http://icg.harvard.edu/ pinyin/, February 1997.

[10] D. Klatt and L. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857, 1990.

[11] K. Kurowski and S. E. Blumstein. Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 76:383–390, 1984.

[12] K. Kurowski and S. E. Blumstein. Acoustic properties for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 81:1917–1927, 1987.

[13] L. S. Larkey, J. Wald, and W. Strange. Perception of synthetic nasal consonants in initial and final syllable position. *Perceptual Psychophysics*, 23:299–312, 1978.

[14] A. Malecot. Acoustic cues for nasal consonants: An experimental study involving a tape-splicing technique. *Language*, 32:274–284, 1956.

[15] D. Recasens. Place cues for nasal consonants with special reference to Catalan. *Journal of the Acoustical Society of America*, 73:1346–1353, 1983.

[16] B. Repp and K. Svastikula. Perception of the [m]-[n] distinction in vc syllables. *Journal of the Acoustical Society of America*, 83:237–247, 1988.

[17] P. F. Seitz, M. M. McCormick, I. M. C. Watson, and R. A. Bladon. Relational spectral features for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 87:351–358, 1990.

[18] R. Sproat, editor. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer Academic Publishers, 1997.

[19] K. N. Stevens. *Acoustic Phonetics.* MIT Press, 2000.

[20] Y. Xu and E. Wang. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33:319–337, 2001.