USING ARTICULATORY ADJUSTMENT TO COMPENSATE FOR HYPERNASALITY — A MODELING STUDY BASED ON MEASURES OF ELECTROMAGNETIC ARTICULOGRAPHY (EMA)

BY

PANYING RONG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Speech and Hearing Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor David Kuehn, Chair
Professor Ryan Shosted, Director of Research
Professor Torrey Loucks
Professor Mark Hasegawa-Johnson
Professor Chilin Shih

Abstract

The speech of individuals with velopharyngeal incompetency (VPI) is characterized by hypernasality, a speech quality related to excessive emission of acoustic energy through the nose, as caused by failure of velopharyngeal closure. As an attempt to reduce hypernasality and, in turn, improve the quality of VPI-related hypernasal speech, this study is dedicated to developing an approach that uses speech-dependent articulatory adjustments to reduce hypernasality caused by excessive velopharyngeal opening. A preliminary study has been done to derive such articulatory adjustments for hypernasal /i/ vowels based on the simulation of an articulatory model (Speech Processing and Synthesis Toolboxes, Childers (2000)). Both nasal /i/ vowels with and without articulatory adjustments were synthesized by the model. Spectral analysis found that nasal acoustic features were attenuated and oral formant structures were restored after articulatory adjustments. In addition, comparisons of perceptual ratings of nasality between the two types of nasal vowels showed the articulatory adjustments generated by the model significantly reduced the perception of nasality for nasal /i/ vowels. Such articulatory adjustments for nasal /i/ have two patterns: 1) a consistent adjustment pattern, which corresponds an expansion at the velopharynx, and 2) some speech-dependent fine-tuning adjustment patterns, including adjustments in the lip area and the upper pharynx.

The long-term goal of this study is to apply this approach of articulatory adjustment as a therapeutic tool in clinical speech treatment to detect and correct the maladaptive articulatory behaviors developed spontaneously by speakers with VPI on individual bases.

This study constructed a speaker-adaptive articulatory model on the basis of the framework of Childers's vocal tract model to simulate articulatory adjustments aiming at compensating for the acoustic outcome caused by velopharyngeal opening and reducing nasality. To construct such a speaker-adaptive articulatory model, (1) an articulatory-acoustic-aerodynamic database was recorded using the articulography and aerodynamic instruments to provide point-wise articulatory data to be fitted into the framework of Childers's standard vocal tract model; (2) the length and transverse dimension of the vocal tract were adjusted to fit individual speaker by minimizing the acoustic discrepancy between the model simulation and the target derived from acoustic signal in the database using the simulated annealing algorithm; (3) the articulatory

space of the model was adjusted to fit individual articulatory features by adapting the movement ranges of all articulators.

With the speaker-adaptive articulatory model, the articulatory configurations of the oral and nasal vowels in the database were simulated and synthesized. Given the acoustic targets derived from the oral vowels in the database, speech-dependent articulatory adjustments were simulated to compensate for the acoustic outcome caused by VPO. The resultant articulatory configurations corresponds to nasal vowels with articulatory adjustment, which were synthesized to serve as the perceptual stimuli for a listening task of nasality rating. The oral and nasal vowels synthesized based on the oral and nasal vowel targets in the database also served as the perceptual stimuli.

The results suggest both acoustic and perceptual effects of the mode-generated articulatory adjustment on the nasal vowels /a/, /i/ and /u/. In terms of acoustics, the articulatory adjustment (1) restores the altered formant structures due to nasal coupling, including shifted formant frequency, attenuated formant intensity and expanded formant bandwidth and (2) attenuates the peaks and zeros caused by nasal resonances. Perceptually, the articulatory adjustment generated by the speaker-adaptive model significantly reduces the perceived nasality for all three vowels (/a/, /i/, /u/). The acoustic and perceptual effects of articulatory adjustment suggest achievement of the acoustic goal of compensating for the acoustic discrepancy caused by VPO and the auditory goal of reducing the perception of nasality. Such a finding is consistent with motor equivalence (Hughes and Abbs, 1976; Maeda, 1990), which enables inter-articulator coordination to compensate for the deviation from the acoustic/auditory goal caused by the shifted position of an articulator.

The articulatory adjustment responsible for the acoustic and perceptual effects as described above was decomposed into a set of empirical orthogonal modes (Story and Titze, 1998). Both gross articulatory patterns and fine-tuning adjustments were found in the principal orthogonal modes, which lead to the acoustic compensation and reduction of nasality. For /a/ and /i/, a direct relationship was found among the acoustic features, nasality, and articulatory adjustment patterns. Specifically, the articulatory adjustments indicated by the principal orthogonal modes of the adjusted nasal /a/ and /i/ were directly correlated with the attenuation of the acoustic cues of nasality (i.e., shifting of F1 and F2 frequencies) and the reduction of nasality rating. For /u/, such a direct relationship among the acoustic features, nasality and articulatory adjustment was not as prominent, suggesting the possibility of additional acoustic correlates of nasality other than F1 and F2.

The findings of this study demonstrate the possibility of using articulatory adjustment to reduce the perception of nasality through model simulation. A speaker-adaptive articulatory model is able to simulate individual-based articulatory adjustment strategies that can be applied in clinical settings to serve as the articulatory targets for correction of the maladaptive articulatory behaviors developed spontaneously by

speakers with hypernasal speech. Such a speaker-adaptive articulatory model provides an intuitive way of articulatory learning and self-training for speakers with VPI to learn appropriate articulatory strategies through model-speaker interaction.

Acknowledgments

I would like to gratefully and sincerely thank my advisors, Dr. David Kuehn and Dr. Ryan Shosted, for their guidance, understanding and patience during my graduate study at the University of Illinois. Their mentorship and encouragement guide me to grow as an independent researcher in the area of Speech and Hearing Science and also as an active collaborator with colleagues from various disciplines. The experience of working with them not only helps me to acquire advanced research skills, but also expands my vision of the long-term career goals.

I would also like to thank my committee, Dr. Torrey Loucks, Dr. Mark Hasegawa-Johnson and Dr. Chilin Shih, who have generously shared their expertise and provided insightful suggestions and continuous support to my study. Their guidance and assistance in the methodological design, experimental access and trouble-shooting of this project have been invaluable. It has been a great pleasure to have the chance of working with all of them in this and a variety of other projects throughout my graduate study at the University of Illinois.

I would also like to thank Dr. Brad Story for generously sharing his dissertation work, which has played a very important role in stimulating my interest in inter-disciplinary study and brainstorming the idea of this project. Additionally, I am grateful to my colleague and collaborator, Christopher Carignan, for assisting me in data collection. I also thank all the participants, who patiently participated in the tedious speaking and listening tasks. Without their cooperation, this project could have not been completed. I am also very grateful to the friendship and support from my colleagues and friends in the Department of Speech and Hearing Science, who have been continuously encouraging me and helping me find participants.

I would like to thank the Department of Speech and Hearing Science, especially the Department Head, Dr. William Stewart, for his assistance in helping me move along with the process of my work, and the department specialist, James Fleener, for his continuous administrative support and help. I would also like to thank the Campus Research Board of the University of Illinois for providing financial support to my study.

Finally, and most importantly, I would like to thank deeply to my parents, Suobao Rong and Wenjin Pan, for supporting me to study abroad with their deepest faith and confidence. Their support, encouragement,

quiet patience and unwavering love have motivated me to overcome many difficulties encountered in my study and life and to continuously challenge myself with great ambition.

Table of Contents

List of Figures

List of Abbreviations

VP	velopharyngeal port
VPO	velopharyngeal opening
VPI	velopharyngeal incompetency
EMA	electromagnetic articulography
EPG	electropalatography
MRI	magnetic resonance imaging
COG	center of gravity
NV	nasal-vowel sequence
MR	magnetic resonance
PCA	principal component analysis
VCV	vowel-consonant-vowel sequence
CVC	consonant-vowel-consonant sequence
CV	consonant-vowel sequence
VC	vowel-consonant sequence
2D	two-dimensional
3D	three-dimensional
EAI	equal appearing interval
DME	direct magnitude estimation
EM	electromagnetic
US	ultrasound
PARAFAC	parallel factors
NM	normalizing map
PONM	parameters orthogonal to the space of NM
O	oral vowel
N	nasal vowel
NA	nasal vowel with articulatory adjustment

TT	tongue tip
TB	tongue blade
ATD	anterior tongue dorsum
PTD	posterior tongue dorsum
UL	upper lip
LL	lower lip
LT	left tragus
RT	right tragus
MT	midpoint of the left and right tragi
ART	articulatory parameters of Childers's model except VPO
LP	linear predictive
LPC	linear predictive coding

List of Symbols

AR1	lower pharyngeal section of Childers's vocal tract model
AR9	upper pharyngeal section of Childers's vocal tract model
AR2	posterior oral cavity of Childers's vocal tract model
AR23	main oral cavity of Childers's vocal tract model
AR4	front oral cavity of Childers's vocal tract model
AR5	lip section of Childers's vocal tract model

Chapter 1

Introduction

Nasalization is the production of a sound while the velum is lowered, so air coming from the subglottal system flows out through both oral and nasal cavities to the atmosphere. Oral vowels produced in conjunction with nasal consonants (/m/, /n/ /ŋ/ in English) tend to be nasalized due to the coarticulatory effect of the velopharyngeal opening (VPO) movement. Previous studies on nasalization focused primarily on: 1) vocal tract and nasal cavity geometry visualized by imaging techniques such as X-ray and magnetic resonance imaging (MRI) (Baer et al., 1991; Dang and Honda, 1994; Moore, 1992; Serrurier and Badin, 2008; Story et al., 1996), 2) speech articulation measured by instruments such as electromagnetic articulography (EMA), electropalatography (EPG), etc., (Aron et al., 2006; Engwall, 2000b, 2001, 2003; Moen and Simonsen, 2007), 3) speech acoustics measured by spectral analysis or computer simulation (Chen, 1997; Feng and Castelli, 1996; Kataoka et al., 2001; Pruthi et al., 2007; Rong and Kuehn, 2010), and 4) speech perception (Beddor and Hawkins, 1990; Ito et al., 2001; Kataoka et al., 2001).

Another type of pathologically-related nasalization occurs in hypernasal speech, which is correlated with velopharyngeal incompetency (VPI), a physical dysfunction caused by anatomical defects such as cleft palate. Although craniofacial surgeries can repair the anatomical defects of VPI, post-surgical behavioral treatment is usually needed to correct the inappropriate articulatory patterns (so-called misarticulation) developed spontaneously by patients with VPI. Currently, such speech therapies rely primarily on speech therapists' subjective perceptual judgement, which has taken into account general principles of hypernasal speech treatment but could not capture the large inter-speaker variability associated with divergent anatomical and articulatory characteristics. As speech is known to be highly variable in terms of both articulation and acoustics across different individuals, traditional therapies for hypernasal speech clearly could not meet the requirement of speaker-customized treatment. Therefore, the long-term goal of this study is to develop a speaker-based therapeutic approach that enables self-training and articulatory learning through speaker-model interaction, aiming at reducing the perception of hypernasality in VPI-related hypernasal speech.

1.1 Geometry of vocal tract and nasal cavity

There have been a considerable amount of studies on vocal tract modeling, many of which are based on two-dimensional imaging, such as X-ray. Volumetric imaging enables direct 3D representation of vocal tract geometry by acquiring a series of image slices, in one or more anatomical planes, through a desired volume of human body. Among many volumetric imaging techniques, Magnetic Resonance Imaging (MRI) can provide three-dimensional static views of a target region with relatively high spatial resolution and minimal risk to human body. Therefore, MRI has been used as a common approach to visualize the anatomical structures of the speech system in a variety of previous studies.

Based on the assumption of one-dimensional wave propagation, the tube vocal tract shape can be approximated as a finite number of cylindrical elements that are stacked consecutively from the larynx to the month. A particular vocal tract shape is superimposed on the tube model by specifying its area function, which is defined as a function of cross-sectional areas of the cylindrical elements with respect to the distance from the glottis. Story et al. (1996) acquired the shapes of the vocal tract during the production of static vowels and consonants by a subject using MRI to image the axial plane of the vocal tract. The images were then analyzed following three steps: 1) segmentation of the airway from the surrounding tissue, 2) 3D reconstruction of the airway by shape-based interpolation, and 3) determination of the airway centerline and measurement of the areas of the oblique sections that are locally perpendicular to the airway centerline. In the first step, a threshold of gray scale value was set up to segment the airway (represented as dark area) from the surrounding tissue (represented as light area). Then a shape-based interpolation technique was used to reconstruct the 3D shape of the vocal tract airway by interpolating segmented image data to form an isotropic (cubic voxel) data set, which served as the binary representation of the vocal tract and was used to compute oblique cross-sectional areas in the next step. Finally, area function was extracted from the interpolated vocal tract shape. This was done by first using an iterative bisection algorithm to compute the centerline and then a voxel counting algorithm to yield the cross-sectional area, anterior-posterior length, and lateral width as a function of slice location along the airway length. These raw area functions were then normalized to a discrete length equal to the speed of sound divided by two times the sampling frequency, as required by the wave-reflection algorithm. Based on the acquired vocal tract area functions, the acoustics of the corresponding speech sounds were simulated and compared to real speech in terms of the lowest three formant frequencies. The results showed a fairly good agreement of formant locations except some centralizing effects in the simulation, which may be attributed to fatigue of articulatory musculature or effects of gravity as the MR imaging is conducted while the subject is in the supine position.

Another way of estimating the geometry of vocal tract is to infer area functions based on 2D midsagittal

profile of vocal tract by converting midsagittal distances to cross-sectional areas. This approach provides an alternative to direct measurement of vocal tract geometry based on volumetric images and is especially useful in dynamic speech imaging, in which the limitation of the temporal resolution of 3D MRI prevents it from acquiring dynamic speech movements without blurring.

Beautemps and P. Badin (1995) used tomography to obtain a set of coherent midsagittal-distance functions for a small corpus of vowels and fricatives produced by one speaker. A speaker-dependent model was then derived and used to convert midsagittal distances to area functions. The conversion was based on Heinz and Stevens' area function model $A = \alpha d^{\beta}$, where A and d are the cross-sectional area and midsagittal distance, respectively, α and β are coefficients depending on speaker characteristics and the location of the section inside the vocal tract. The acoustic characteristics of the corresponding speech sound was simulated based on the estimated area function. Comparisons with the acoustics of real speech showed a general agreement on formant frequencies, suggesting the reliability of the area function estimation model.

Beautemps and P. Badin (1995) further modified the $\alpha - \beta$ model to improve its reliability at both physiological and acoustic levels. Midsagittal function was first defined as the distance between the upper and lower contours of the midsagittal vocal tract profile, measured on the line perpendicular to the midline of the vocal tract, as a function of the abscissa in this midline. Midline was derived as the line for which the wave front at any point is always perpendicular to the tangent to the line at this point. The length of each section was computed as the sum of the length of the two straight segments of the midline located within the section. Then midsagittal distance was calculated as the ratio of the surface of the section over its length. With the midsagittal distance function, the $\alpha - \beta$ model was modified to include two sets of Fourier series coefficients, which controlled $\alpha_{inf}(x)$ and $\alpha_{sup}(x)$, respectively. These coefficients were optimized for the whole corpus (three vowels and six fricatives) to minimize the quadratic error of pressure at frequencies when pressure node occurs at the lip end of the vocal tract. The simulated area functions based on the optimized coefficients indicated non-monotonicity between midsagittal distance and area occurred mainly in the front regions of the vocal tract, but never happened at the lips and very rarely in the region around the incisors. Comparison of formant frequencies between the simulated results and real speech revealed average errors of 8% for F1, 3.1% for F2 and 1.6% for F3. The higher errors related to F1 could be explained partly by relatively higher measurement errors. The results suggested the area functions derived from the midsagittal-distance-to-area conversion model generated outputs with acoustic characteristics similar to the signals recorded simultaneously with radiographs. The consistency between the simulation and real speech could be attributed to the highly-constrained mode of the midsagittal-distance-to-area conversion, which derived the coefficients based on a variety of sounds including both vowels and consonants to ensure the reliability of the model.

In order to study nasal sounds, the nasal tract is modeled as a side branch coupled to the main vocal tract at the opening of the velopharyngeal port (VP). Anatomically, the nasal tract differs from the oropharyngeal cavity in its 1) static structure, 2) complexity of geometry including asymmetry between left and right nasal passages, which may introduce extra spectral poles and zeros acoustically, 3) covering of mucosal layers, which causes extra energy damping, and 4) existence of paranasal cavities, which result in more spectral complexities. Dang and Honda (1994) used MRI to acquire the 3D geometry of the nasal and paranasal cavities by imaging the nasal tract in the coronal orientation and to obtain the geometry of the oropharyngeal cavity by imaging in the sagittal orientation during production of sustained nasal consonants. Audio recording was conducted in a separate experiment with the same speech materials. The MRI images were processed in the following three steps: 1) reconstructing 3D profiles of the nasal and paranasal cavities using a software called VoxelView; 2) detecting air-tissue boundary at a desired threshold value using a general purpose image processing software Photoshop; and 3) obtaining the area and circumference of the airway in each section by counting the numbers of the pixels inside and along the airway boundaries. An assumption of plane wave propagation was made, which assumes the propagation of wave inside the oropharyngeal and nasal tracts is only dependent on the area function of the corresponding tract. Based on the imaging results, Dang and Honda (1994) suggested dividing the nasal tract into the posterior, middle, and anterior portions, each of which differs in length, volume, and morphology. The average length of the section was calculated to be 3.9 cm for the posterior portion, 4.4 cm for the middle portion, and 3.4 cm for the anterior portion. With regard to the complexity of the nasal tract geometry, a shape factor was defined as $\frac{S}{4\pi A}$ to account for the geometrical differences among the three portions of the nasal tract, where S and A are the cross-sectional circumference and area of the nasal tract, respectively. The shape factor was determined to be of the order of 1 for the posterior portion, 4 for the middle portion, and 2 for the anterior portion. To account for the asymmetry of the nasal tract, an asymmetry coefficient was introduced to the model. Paranasal cavities (i.e., sphenoidal sinus (SS) and maxillary sinus (MS)) were simplified and modeled as Helmholtz resonators coupled to the nasal passages as side branches, which contribute to the formation of extra spectral pole-zero pairs. Given the mobility of the velum, it was excluded from the static imaging view of the nasal tract and, instead, was included as part of the main vocal tract in a sagittal view. Finally, cross-sectional areas of the entire vocal tract were derived from the coronal view of the nasal tract and the sagittal view of the oropharyngeal cavity.

To examine the effects of asymmetry of the nasal tract and coupling of the paranasal cavities on the acoustics of nasal speech sounds, Dang and Honda (1994) developed a dual-tube model based on the structural information derived from MR images. In the nasal tract model, paranasal cavities were coupled to the nasal passages as discrete Helmholtz resonators without end correction. The acoustic outcome suggested both

asymmetry of the nasal tract and resonances of the paranasal cavities introduced extra poles and zeros in the transfer function of the nasal sounds. Comparisons between the transfer function simulated by the model and the spectral envelop measured from real speech signals suggested a fairly good agreement on the overall spectral pattern below 2kHz.

As the geometry and location of paranasal cavities vary across different individuals, plus coupling between the oropharyngeal cavity and nasal cavity depends on the position of the velum, the anti-resonances dependent on the resonating characteristics of the paranasal cavities should vary in terms of frequency and amplitude across different speakers and speech tokens as a result. In order to represent the acoustic outcome caused by the resonances of the paranasal cavities, a four-zero model was constructed to characterize the anti-resonances caused by paranasal sinuses (Dang and Honda, 1996). A special approach was designed to distinguish the anti-resonances introduced by different paranasal cavities. Based on the fact that the transmission characteristics related to spectral zeros vary only when the input stimulus passes through the branch location, an acoustic recording was carried out inside the nasal tract using a probe microphone with a 30-cm-long flexible tube, where the location of the tip of the probe tube was moved from the anterior nostrils to the posterior choanae, in 0.5-cm intervals. The speech stimuli consisted of two sustained nasal consonants (/m/ and /n/) and ten NV tokens. The transmission characteristics of each segment of the nasal tract were measured by averaging the spectral envelopes across all samples of the nasal consonants /m/ and /n/, where the nasal tract geometry was assumed to stay static across these consonant samples. The anti-resonant frequency of each paranasal cavity was determined following a two-step procedure: 1) estimation of the ostium locations based on the positions of the zeros in the spectra of a series of measurements, 2) identification of the anti-resonances and their affiliated paranasal sinuses by matching the estimated ostium locations with the real locations of the ostia (Dang and Honda, 1996). Due to different locations of the paranasal (i.e., sphenoidal, maxillary and frontal) sinuses, the anti-resonances affiliated with the frontal, maxillary and sphenoidal sinuses appeared sequentially in the output spectrum when the corresponding sinuses were incorporated into the system as the probe tube was moved from the anterior to the posterior nasal tract. The anti-resonant frequencies were determined to range from 750 Hz to 1900 Hz for the sphenoidal sinus, from 400 Hz to 1100 Hz for the maxillary sinuses, and from 630 Hz to 2060 Hz for the frontal sinuses. Strong left-right asymmetries and large individual variations were observed based on the measurement of anti-resonant frequencies.

The geometrical information of the oropharyngeal and nasal tracts enables development of a transmission line model to simulate the transfer function of the vocal tract. From an acoustic view, the distribution of the anti-resonances uncovered the acoustic effects of coupling of the paranasal cavities, which were, to introduce two spectral zeros below 1 kHz and two zeros above 1 kHz, and therefore, a "four-zero model."

In a transmission line model, coupling between the oropharyngeal tract and the nasal cavities is realized through a three-port connection, namely, the velopharyngeal port. Serrurier and Badin (2008) developed a 3D linear articulatory model of the velum and the nasopharyngeal wall based on MRI and computed tomography (CT) images of a French subject sustaining a set of 46 speech sounds, covering his articulatory repertoire. Principal component analysis (PCA) was applied to the images at the VP, so that the first principal component was found to relate to a head tilt movement in the sagittal plane, but its contribution was removed from the data due to lack of relevance to speech. After the correction for head movements, PCA uncovered another two main degrees of freedom for velar movement. A dominant component accounting for 83% of the variance of velar movement corresponded to an oblique vertical movement related to the pulling action that was expected from the activity of the levator veli palatini muscle. The second component accounting for 6% of the variance of velar movement represented a mostly horizontal movement that corresponded rather well to the sphincter-like action plausible for the activity of the superior pharyngeal constrictor muscle. It was also shown that the movement of the nasopharyngeal wall was very much related to the first component, which accounted for 47% of the variance of its movement. Therefore, the VP model was constructed based on two controlling articulatory parameters corresponding to the two PCA components, respectively. Adding the VP model to the 3D model of the main vocal tract resulted in a 3D geometric representation of the entire vocal tract, which was used to simulate the acoustic characteristics of different vowels. The simulated results showed a reduction of the F1-F2 vowel space in the 450-1000 Hz zone when the shape of the vocal tract evolved from pure oral vowels to nasal sounds that transmitted through nasopharyngeal tract (Feng and Castelli, 1996). The simulated acoustic results in Serrurier and Badin (2008) generally agreed with Feng and Castelli (1996). On the other hand, the simulation of the 3D generic mesh of the velum was consistent with the position of a freshpoint on the lower surface of the velum tracked by an EMA sensor when the same subject was recorded to produce a comprehensive set of vowel-oral consonant-vowel (VCV) sequences.

1.2 Acoustics of nasal sounds

According to the previous discussion, MRI has been used for volumetric imaging of the vocal tract during production of sustained speech sounds in recent studies (Alwan et al., 1997; Baer et al., 1991; Dang and Honda, 1994; Matsumura, 1992; Moore, 1992; Narayanan et al., 1995, 1997). The geometry of vocal tract derived from MRI has been used to model the vocal tract and simulate the acoustic output. Based on the area functions in Story (1995); Story et al. (1996), Pruthi et al. (2007) used a computer vocal tract model, VTAR (Zhang and Espy-Wilson, 2003), to simulate the spectral properties of nasal vowels. Two modeling methods of oral-nasal coupling were compared: 1) trapdoor coupling method, in which the area of the first

section of the velopharynx was set to the desired coupling area and no other changes were made to the areas of either the velopharynx or the oral cavity, 2) distributed coupling method, in which not only the first section of the velopharynx was set to a desired coupling area, but also the rest of the velopharynx were linearly interpolated to get a smooth variation in areas, resulting in a corresponding reduction of the area inside the oral cavity in the velopharyngeal region. It was found that, in the trap door coupling method, the zeros in the simulated transfer functions did not change with the nasal coupling area and, therefore, the order of the principal cavities that the poles are affiliated with was fixed. Contrastively, the distributed coupling method performed in a more realistic way such that the order of the principal cavities that the poles are affiliated with changed with the coupling area. More advantages of the distributed coupling method was found in Maeda (1982b), which suggested reduction of the area inside the oral cavity is important to produce natural sounding nasal vowels. Furthermore, Serrurier and Badin (2008) examined the influence of velar movement on the area function based on the simulation of their 3D VP model and found a covariation of oral and nasal areas when velum moved.

Pruthi et al. (2007) examined the influence of the following factors on the acoustics of nasal vowels: 1) degree of oral-nasal coupling, 2) asymmetry between the two parallel nasal passages, and 3) coupling of maxillary and sphenoidal sinuses (frontal and ethmoidal sinuses were discarded due to the lack of structural data). Based on a vocal tract model, two simulations were done, including: 1) transfer function, which examines the spectral characteristics of nasal vowels and 2) susceptance plot, which uncovers the evolution pattern of spectral poles and zeros. The pole-zero pattern was determined by the crossings of two curves, $-(B_p + B_o)$ and B_n, where B_p, B_o, and B_n stood for the susceptance of the pharyngeal, oral, and nasal cavities, respectively, looking from the coupling location to the corresponding cavity. Here evolution of the spectral pole-zero pattern is defined as the change of spectral pattern (e.g. extra poles and zeros, shift of formant frequencies, change of formant affiliations, etc.) caused by the variation of oral-nasal coupling degree. Susceptance plots informed the order of principal cavity affiliation for each pole in the spectrum of a dual-branched system.

A simplified distributed system corresponding to a lossless transmission line model was used to compute susceptance plots and transfer functions in Zhang and Espy-Wilson (2003). The effect of oral-nasal coupling was examined in the evolution pattern of spectral poles and zeros. It was found that the frequency of a pole was not necessary to increase monotonically with increasing of the coupling area due to a possible change in principle cavity affiliation. On the other hand, with an increase of coupling degree, the paired nasal pole-zero became more split, resulting in a more distinctive nasal pole and an attenuated oral formant due to the canceling effect of the adjacent nasal zero. The effect of asymmetry of nasal passages was explored by comparing a one-tube model, which combines the left and right nasal passages as a single tube with its

cross-sectional area equal to the sum of two passages, with a two-tube model, which simulated the left and right passages as two parallel tubes. Additional pole-zero pairs at around 1649Hz and 3977Hz were found in the two-tube model of nasalized /a/ with an oral-nasal coupling area of $0.4cm^2$, suggesting asymmetry of the two nasal passages as the main cause of these extra poles and zeros. This is because one passage acts as a zero-impedance shunt at a particular frequency to short-circuit the other passage, resulting in a zero at this frequency in the output spectrum. Coupling of paranasal sinuses introduced extra spectral poles and zeros as well. In Pruthi et al. (2007), maxillary sinuses (MS) and sphenoidal sinuses (SS) were modeled as two Helmholtz resonators connected to the nasal passages, where the resonances of the sinuses appeared in the simulated transfer functions as spectral zeros. Despite the static structures of the sinuses, the frequencies of the spectral zeros related to the resonances of the sinuses could vary when the oropharyngeal configuration changed, as the output spectrum was computed based on the acoustic signal transmitted to the air as a combination of both oral and nasal sound pressures. It was found that the spectral zeros affiliated to MS were in the range of 620-749 Hz, and the zeros affiliated to SS stayed in the range of 1527-1745 Hz.

Pruthi et al. (2007) summarized the acoustic characteristics of nasal vowels as: 1) extra poles and zeros in the spectrum, which are caused by coupling of nasal cavities, asymmetry of nasal passages and branching of paranasal sinuses coupled to the nasal passages, 2) reduction of F1 amplitude, caused by presence of the low-frequency anti-resonances, 3) increase of the bandwidths of all poles affiliated to the nasal cavity, 4) flatness of low-frequency spectral envelope, 5) movement of low-frequency center of gravity (COG) towards a neutral vowel configuration (i.e., schwa), 6) reduction of overall intensity due to presence of nasal zeros, and 7) shift of pole frequencies due to switch of affiliations from the oral cavity to the nasal cavity.

As production of nasal vowels depends not only on the anatomy of the speaker's nasal cavity, but also on the superimposed vowel configuration and the degree of oral-nasal coupling, these complexities should all be reflected in the spectra of nasal vowels as unique acoustic characteristics. Maeda (1982a) proposed global flattening of spectra in the F1-F2 region (around 200–2000 Hz) as an independent cue of nasalization and two spectral peaks (at 250 Hz and 1000 Hz) as the most evident correlates of the perceived nasality. It was also observed that the distance between these two nasal peaks were strongly correlated with the degree of nasal coupling. Hawkins and Stevens (1985) suggested presence of a low-frequency peak and its wide bandwidth, taken together with a pole-zero pair caused by the coupling between the oral and nasal cavities, played a fundamental role in the perception of nasality. These studies summarized the primary acoustic cues of nasalization, but lack of consistency in the spectral features of speech across different individuals makes it difficult to generalize these conclusions.

In order to examine how the variation of vocal tract configuration affects the acoustics of nasal vowels, Feng and Castelli (1996) simulated a nasal " target," which corresponds to the configuration of the vocal

tract when the velum is completely lowered to shut the oral cavity from the nasopharyngeal tract. With this target, the shape of vocal tract for a nasal vowel was simulated as an intermediate configuration during the evolution from an oral configuration to the simulated nasal target configuration. As phonemic nasal vowels were considered to originate from assimilation of oral vowels to adjacent nasal consonants (Ferguson, 1963, 1975), Feng and Castelli (1996) correspondingly introduced a /ŋ/-like nasal-consonantal target that the shape of the vocal tract tends to approach during nasalization. Anatomically, the shape of this /ŋ/-like nasal target corresponded to the nasopharyngeal tract, which is divided into a nasal section and a pharyngeal section. The pharyngeal section was modeled based on the images of the vocal tract during oral vowel productions, so there were slight variations across different vowels. The nasal section, on the other hand, maintained a stable configuration for all vowels produced by a specific speaker. Due to the configurational differences in the pharyngeal section for different vowels, the lowest two resonant frequencies of the nasopharyngeal tract were not fixed, but rather occupied a small region in the F1-F2 plane, centered at around 300–1000 Hz. In this way, the nasal target provided a virtual anatomical structure, the shape of which was slightly sensitive to vowel contexts, but much more stable compared to the speech-dependent variation of the oropharyngeal configuration Feng and Castelli (1996).

As the nasopharyngeal target is realized through coarticulation of the velar nasal consonant /ŋ/ with an oral vowel, speakers were asked to simulate this process by first producing an oral vowel without phonation, and then lowering the velum completely, while moving the other articulators as little as possible (Feng and Castelli, 1996). Then productions of nasal vowels were simulated as intermediate states between the corresponding oral vowel configurations and the nasopharyngeal target. Their transfer functions showed two principal patterns: “simple and “complex. The “simple” pattern of transfer function presented four to six peaks, whereas the “complex” pattern showed eight to ten peaks. The peaks of a “simple function were principally located in the following frequency ranges: 250–300 Hz, 650–700 Hz, 900–1100 Hz, 1900–2200 Hz, 2800–3300 Hz, and around 4000 Hz, whereas a zero at around 500–600 Hz appeared in some cases. The frequency ranges of the peaks in a complex function included: 250–300 Hz, 400–450 Hz, 650–700 Hz, 850–900 Hz, 1200–1300 Hz, 1800–2200 Hz, 2800–3300 Hz, and around 4000 Hz, whereas in addition, a zero appeared at around 1200 Hz. In some cases, extra zeros also appeared between F1 and F2, F2 and F3, and F3 and F4, respectively. Based on the patterns of the simulated transfer functions, Feng and Castelli (1996) concluded that F5, F8, F9, and F10 were caused by the resonances of the nasopharyngeal tract, whereas F2, F3, and F4 might have their origins in the resonances of sinus cavities, and F6 and F7 were attributed to the asymmetry of the left and right nasal passages. The low-frequency nasal resonances are regarded as the acoustic correlates of the perceived nasality, whereas the higher-frequency nasal resonances introduce extra poles to increase the high-frequency spectral intensity. As the low-frequency spectral intensity is attenuated

by coupling of the nasal tract and the high-frequency spectral envelope is modulated by extra nasal poles, the overall spectral center-of-gravity could be shifted as a result, which might change the perceptual speech quality (Ito et al., 2001).

As the shape of a nasal vowel was considered as an intermediate state when the vocal tract evolves from an oral vowel configuration to the /ŋ/-like nasal target, the evolution of transfer functions from the oral vowel to the corresponding nasal target uncovers the relationship between the degree of nasal coupling and the relevant acoustic characteristics (Feng and Castelli, 1996). Due to lack of sufficient structural data on the velopharynx, an assumption was made to divide the velopharyngeal coupling port into two parts: a nasal port connected to the nasal cavity and an oral port serving as the entrance of the oral cavity. In the models of nasal vowels, two parallel L-C (-R) electrical circuits were used to simulate oral and nasal branches, whereas the degree of oral-nasal coupling was controlled by multiplying one branch by an access-coefficient d_c and multiplying the other branch by the reciprocal $1/d_c$. Among all simulated transfer functions, seven evolution patterns were discovered, which were distinguished by the distributions of resonant and anti-resonant frequencies. Eleven French vowels were categorized into three groups, which were represented by the three corner vowels /i/, /u/, and /a/. The relevant acoustic characteristics for these three vowel groups included: 1) acquisition of a high-frequency pole (1000 Hz) and disappearance of F2 for the /i/ category; 2) lowering of F1 frequency and/or elevation of F2 frequency for the /u/ category; and 3) acquisition of a low-frequency pole (250–300 Hz), attenuation of F1 and lowering of F2 frequency for the /a/ category. As these simulations were conducted based on a single-tube model of the nasal tract, the effects of sinus coupling and asymmetry of nasal passages were not accounted for. To take into account the role of paranasal sinuses in reshaping the low-frequency spectrum, maxillary sinuses were added to the original single-tube model as a side branch of the nasal cavity, which introduced an extra pole-zero pair in the transfer function, whereas the overall pattern of pole evolution stayed stable.

The spectral characteristics of three French nasal vowels (/ĩ/, /ã/, /ũ/) were analyzed from real speech recordings and compared with the simulated results. A "complex" pattern similar to that derived from the simulation of the nasopharyngeal target was found in the spectra of these French nasal vowels. The distance between the first spectral peak and the nasal peak at around 1000 Hz was found to serve as the essential cue that distinguished different nasal vowels. Therefore, the nasopharyngeal target proposed by Feng and Castelli (1996) serves as a reliable virtual target for modeling of nasal vowel productions.

1.3 Perception of nasal sounds:

The primary questions related to the perception of nasal speech are: 1) how the perception of nasality correlate with the characteristic acoustic cues; 2) how to quantify the perception of nasality. For the first question, previous studies have examined the relationship between the acoustic and perceptual correlates of nasal sounds. Hawkins and Stevens (1985) suggested the acoustic cue of nasalization was the amplitude of the peak in the F1 region of the spectrum. Chen (1997) introduced two acoustic parameters, $A1 - P1$ and $A1 - P0$ to represent the perceptual correlates of nasalization, where $A1$, $P0$, $P1$ were the amplitudes of the first oral formant (F1) and the first and second nasal peaks in the vicinity of F1, respectively. These findings were based on formant analyses, which used individual formant values to serve as the acoustic cues of nasality. However, nasal vowels are in general, characterized by broad peaks and flat spectral envelop so that individual formants are usually not as prominent as those in the the oral vowel spectra. Therefore, some other studies took into account the shape of the whole spectrum to uncover the acoustic correlates of nasality.

Kataoka et al. (2001) suggested the 1/3-octave analysis as an alternative to examine the acoustic characteristics of hypernasality in children. This bandwidth was selected because it is comparable to the critical bandwidth of the perceptual mechanism of the hearing system. The study compared the average 1/3-octave spectra of /i/ between a group of hypernasal speakers and a group of speakers with normal resonance. Based on the comparison, the acoustic correlates of hypernasality were identified as increased amplitudes between F1 and F2 and decreased amplitudes around F2. These acoustic features were manipulated to generate 36 speech samples to serve as the speech stimuli for a perceptual experiment, in which the severity of hypernasality was evaluated by four experienced listeners using 6-point equal-appearing interval scaling method. The relationship between the acoustic parameters and the perceptual ratings of hypernasality from the four listeners was examined using a multiple regression analysis. The nasality scores estimated from the regression equation ($y = 7.42 + 0.11 \cdot L_{1k} + 0.06 \cdot L_{1.6k} - 0.17 \cdot L_{2.5k}$, where L_{1k}=level of the 1kHz band, $L_{1.6k}$=level of the 1.6kHz band, and $L_{2.5k}$=level of the 2.5kHz band) given the amplitudes of three 1/3-octave bands (i.e., 1kHz, 1.6kHz, and 2.5 kHz) were found to be highly correlated ($R = 0.84$) with the perceptual ratings of nasality by the listeners. Comparisons of the standardized regression coefficients across different listeners found almost identical coefficients for L_{1k} and $L_{1.6k}$, but considerably different coefficients for $L_{2.5k}$, suggesting that listeners placed nearly the same weight on the levels of 1kHz and 1.6kHz, but different weights on the level of 2.5kHz. In order to examine the influence of different frequency components on the perception of hypernasality, the speech samples were divided into four groups based on the degree of manipulation on different frequency components. The results suggested: 1) no obvious change was observed in the nasality

ratings when the amplitude of the 1kHz band was increased by more than 15 dB, or when the amplitude of the 1.6 kHz band was increased by less than 10 dB, 2) decreasing the amplitude of the F2-F3 band resulted in increased ratings of hypernasality, 3) ratings of hypernasality increased as the amplitude of F1 increased, but remained stable when the amplitude of F1 decreased and 4) modifying the amplitude of F0 in either direction increased ratings of hypernasality.

The 1/3-octave analysis developed by Kataoka et al. (2001) overcomes the drawback of formant analysis by considering the entire low-frequency spectral envelope rather than individual formants. One limitation lies in that changes of spectral envelope in the F2-F3 frequency range shift the overall spectral balance, resulting in change of the perceived speech quality, which might influence the perceived degree of nasal resonance when spectral peaks are not prominent (Hawkins and Stevens, 1985). Changes of speech quality and severity of hypernasality could both influence listeners' rating of nasality scores and might result in larger inter-listener perceptual variablity. In this sense, it is important to take into account both speech quality and nasality when considering the perceptual properties of nasal/nasalized vowels.

Although it is a well-accepted statement that the lowest two or three formants are responsible for the perception of vowel quality, there have also been studies suggesting that individual formants are not exclusive cues but the whole spectral shape is also responsible for the perception of vowel quality. Ito et al. (2001) conducted three perceptual experiments to explore the perceptual correlates of vowel quality. In the first experiment, the control stimuli were synthesized using a cascade-type Klatt synthesizer (Klatt, 1980). The first or second formant of the stimuli was suppressed as much as possible while maintaining the original spectral shape. If the vowels were perceptually recognized on the basis of F1 and F2, then the perception of vowel quality would not vary when the corresponding formant is suppressed. However, it was found that the phoneme boundaries for F1-suppressed stimuli were closer to the control stimuli than expected, whereas the similarity between F2-suppressed stimuli and the control stimuli was even larger except in the region of $F1 = 600Hz, F2 = 1200Hz$. The results indicated that vowel quality, especially the quality related to the place of articulation (front/back), could be perceptually identified even if the F2 information was not available, so there should be other cues responsible for the perception of place of articulation in addition to the F2 frequency.

Ito et al. (2001) suggested the amplitude ratio of the high-to-low frequency components could be one of the potential cues of the perceived vowel quality. To test this hypothesis, a second experiment used 81 perceptual stimuli synthesized with different spectral envelopes focusing on various F1 and A1, where F1 varied between 250 and 1250 Hz in 125-Hz steps and the amplitude difference between the first and second formants (i.e., $A1 - A2$) varied between 0 and 48 dB in 6-dB steps. It was observed that changes of $A1 - A2$ could lead to changes of the perceived vowel quality. In this sense, the amplitude ratio of the high-to-low frequency

components could be an essential cue responsible for the perception of place of articulation. However, it is not clear whether the amplitude ratio still plays the same role when the formant structure information is available. In order to compare the influence of amplitude ratio and individual formants in determining the perception of vowel quality, a third experiment was conducted with the vowel stimuli synthesized based on five formants, where F2 was at a constant frequency of 1500 Hz and the amplitude ratios $A1/A3$ and $A4/A5$ were variable. The results suggested the variation of amplitude ratio ($A1/A3$, $A4/A5$) could still change the perceptual results under the influence of fixed formant frequencies. The perceived phoneme boundaries were quite similar to the boundaries of the control stimuli in Experiment 1. Therefore, a nominal F2 was defined based on the amplitude ratio of the high-to-low frequency components to serve as the acoustic cue correlated with the perceived place of articulation. Specifically, stimuli with high nominal-F2 were perceived as front vowels and those with low nominal-F2 were perceived as back vowels.

The study of Ito et al. (2001) demonstrated that formant frequencies were not exclusive cues that determine vowel perception, and the amplitude ratio of the high-to-low frequency components somehow plays a more important role in determining the perceived place of articulation. As a change of vowel quality can influence the perception of nasality, it is necessary to examine how the change of spectral shape caused by nasalization correlates with the perception of vowel quality and nasality.

In addition to place of articulation, vowel height is another perceptual factor that would influence the perception of nasality. When the first and second formants of a vowel were separated by less than 3.5 Bark, the perception of vowel height and other relevant qualities are determined by the weighted average of the low-frequency spectral envelope (center of gravity) rather than by individual harmonics or formants. Beddor and Hawkins (1990) suggested that when two or more spectral peaks lay within 3.5 Bark of one another, F1 and its centroid (an amplitude-weighted average of frequency) roughly determined the boundaries within which the perceptual COG of vowel height lay. Therefore, although the frequencies of formants dominate perceptual responses when formant bandwidths are narrow, the overall spectral envelope leaves more influence when formant bandwidths are wide.

The lowest spectral peak of a nasal vowel usually has a smaller amplitude and wider bandwidth than non-nasal peaks. Measurements on naturally spoken nasal vowels from a number of languages suggested the shifted F1 and the lowest nasal formant are usually within 3.5 Bark of one another (Beddor, 1983), so the overall spectral envelope should be more influential in determining the quality of nasal vowels according to Beddor and Hawkins (1990). In order to examine the effect of center-of-gravity on vowel height, Beddor and Hawkins (1990) compared the perceived vowel height between multi-formant synthetic nasal and oral vowels. The frequency of COG was set to range from 100 Hz to 1100 Hz for all vowels except the /a/ series, for which the upper frequency bound was extended to 1400 Hz to accommodate all low-frequency peaks and

skirts, including F1 in both oral and nasal vowels, FN (the first nasal formant) in nasal vowels, and F2 in back vowels. It was found the oral vowels were perceptually most similar to non-high nasal reference vowels when their F1 and centroid frequency fell between the centroid and F1 frequencies of the nasal reference vowels. The responses were skewed toward the centroid frequency of the nasal reference vowel, suggesting the centroid frequency is more influential in determining vowel quality than F1 in non-high vowels (i.e., /e/, /o/, /ae/, /a/). The only exception was nasal /i/ (with an F1-FN spacing as large as 4.5 Bark), which has over 70% responses determined by F1-matches between oral and nasal vowel counterparts. Therefore, in the low-frequency spectrum, the frequencies of well-defined spectral peaks and the overall spectral envelope around less prominent peaks both play important roles in determining the perception of vowel height.

To further investigate the relation between spectral characteristics and perceptual COG, a second experiment was conducted to examine the effect of formant bandwidth, fundamental frequency and duration on vowel perception using two-formant and one-formant synthetic vowels (Beddor and Hawkins, 1990). The results suggested the perceived quality of two-formant vowels corresponded to neither F1-match nor centroid-matched one-formant vowels. However, the matches were significantly closer to F1 than the centroid frequency for vowels /o/ and /u/. Comparison of the results between experiment 1 and experiment 2 suggested different dominant factors determinant to vowel perception. Such difference might be attributed to the influence of high-frequency spectral components of the multi-formant vowels and the complex low-frequency spectral features of nasalization characterized by broad peaks separated by shallow troughs.

To test the hypothesis that a broad, flat spectral prominence leads to a higher perceptual COG, two-formant vowels with different degrees of formant prominence and manipulable bandwidth were synthesized to serve as stimuli for experiment 3 (Beddor and Hawkins, 1990). In this experiment, narrow-banded and wide-banded references were compared with medium-banded stimuli. The results showed significant different numbers of responses on centroid-matched narrow-banded /o/ across listeners but no significant inter-listener difference for narrow-banded /a/, whereas wide- and medium-banded vowels with the same formant and centroid frequencies were not selected as the best match with the reference vowels. The centroid frequency of the best-match medium-banded stimulus was 30 Hz higher than that of the wide-banded reference vowel for the /a/ series, but the frequency shift was in the opposite direction for /o/. Such bidirectional shift was explained as the effect of spectral shape (i.e., formant bandwidth, prominence, etc.) on the perceived vowel quality.

According to Beddor and Hawkins (1990), exclusive peak-picking methods are only appropriate for vowels with well-defined spectral peaks, such as front oral vowels, to determine the perceptual correlates. Nasal vowels, characterized by broad and attenuated low-frequency formants, rely more on the overall spectral shape to determine the perceptual properties. Back vowels represent an intermediate case in terms of

peak prominence, compared to front vowels and nasal vowels. In this sense, a comprehensive model of vowel perception (particularly the perceived vowel height) should follow a mechanism that produce different auditory responses dependent upon variable prominence/bandwidth of formant peaks. When spectral peaks are well-defined, the peak frequencies dominantly determine the perceptual response, whereas the overall spectral shape becomes more influential in determining the perceptual quality of vowels with less well-defined formants.

So far, the relationship between the spectral characteristics and the perceptual correlates of nasal vowels has been discussed. To determine how much the perceived nasality changes with its acoustic correlates, there needs to be a way of quantifying the degree of nasality.

The quantification of nasality requires an appropriate psychophysical scaling method. Equal-appearing interval (EAI) scaling and direct magnitude estimation (DME) are two commonly-used methods of voice quality scaling. Zraick and Liss (2000) compared the effectiveness of these two methods in rating of nasal voice quality and suggested DME served as a more appropriate method for nasality rating.

In DME, listeners scale individual speech samples relative to each other or to a standard stimulus (modulus), which is usually obtained along the mid-range of the perceptual dimension. A numerical value (for example, 100) is assigned to the modulus so that listeners can scale the respective perceptual attribute of the stimulus relative to the modulus. The advantages of DME in perceptual rating include 1) DME does not assume a linear partition of the continuum (Schiavetti et al., 1981) and 2) DME is not bounded by fixed minimum/maximum values (Stevens, 1975). Before comparing EAI and DME scalings, the nature of perceptual dimension should be discussed first. Stevens (1975) described two classes of dimensions, namely, metathetic and prothetic. A metathetic dimension is one that varies in terms of quality and sometimes described as substitutive. A prothetic dimension, on the other hand, is one that varies in terms of quantity or magnitude, and is therefore described as additive. Stevens (1974) showed that a prothetic continuum is not amenable to linear partitioning, so EAI is not appropriate to quantify prothetic continuum. For a metathetic dimension, however, Stevens (1974) showed that listeners were able to divide the continuum into equal intervals, that is, listeners' naturally occurring perceptual intervals are equal. Stevens (1975) outlined a procedure, which plots mean EAI scores against mean DME scores to determine whether a perceptual dimension falls along a metathetic or prothetic continuum based on the relationship between the two. Specifically, a linear relationship indicates that listeners assign equal perceptual spaces to the intervals of the EAI scale, suggesting a metathetic continuum, whereas a nonlinear relationship suggests a prothetic continuum.

Zraick and Liss (2000) used a Klatt formant synthesizer to synthesize the oral vowel /i/ at five fundamental frequencies (80 Hz, 120 Hz, 180 Hz, 220 Hz, and 300 Hz), each of which was sustained for 1.5 seconds. Then four different nasalized cohorts that simulated a continuum of nasality from mildly nasal to severely nasal

were synthesized corresponding to each oral vowel. With a total of 25 stimuli, listeners were instructed to rate the nasality of the vowel stimuli 1) on a 5-point EAI scale, with a rating of 1 representing least nasal and a rating of 5 corresponding to most nasal, and 2) using the DME method, with the standard stimulus (i.e., modulus) selected from the mid-range of the continuum (corresponding to Point 3 on the 5-point EAI scale of nasality). Then the linearity of the relationship between the mean EAI ratings and the mean DME ratings was evaluated through a simple linear regression analysis. Besides, the internal consistency of rating and inter-rater reliability were assessed for each scaling method by calculating the coefficient alpha and the intraclass correlation coefficient, respectively. The results suggested the internal consistency of nasality ratings determined by DME was extremely good ($> .9$), whereas the internal consistency determined by EAI was extremely poor ($< .2$). On the other hand, the inter-judger reliability for the DME method was considerably better than that the EAI method. The comparison between EAI and DME nasality ratings revealed a significant curvilinear relationship, suggesting a prothetic rather than a metathetic continuum better represents the perceived nasality. Therefore, DME is a more appropriate scaling method to quantify nasality than EAI.

Zraick and Liss (2000) summarized a number of limitations of EAI scaling in quantifying nasality, including: 1) the bias exhibited by listeners when attempting to partition nasality into equal intervals, 2) failure of EAI to capture a responder's full range of perception scale (Stevens, 1975) and 3) the tendency that listeners try to divide stimuli into categories so that all categories are used equally often in EAI (Gescheider, 1976). These issues limit the application of EAI in rating of nasality, whereas the DME method proves to be more appropriate.

With regard to the perception of nasal vowels, an important consideration is whether listeners' perception of vowel quality interacts with the perception of nasality. Previous studies showed that the perception of the quality of normal speech is multidimensional, that is, a listeners perception of speech quality involves integrative judgments of more than one dimension (Colton, 1987; Matsumoto et al., 1973; Singh and Murry, 1978; Walden et al., 1978). In nasal vowels, many acoustic features other than those related to nasalization were suggested to have potential influences on speech intelligibility (Kent et al., 1989; Phillips and Kent, 1984). On the other hand, listeners' rating of nasality was regarded to be lack of reliability (Bradford et al., 1964; Counihan and Cullinan, 1970; Fletcher, 1976) due to the influence of a number of segmental and suprasegmental features (Bzoch, 1989; Carney and Sherman, 1971; Counihan and Cullinan, 1972; Moore and Sommers, 1973). Therefore, it is reasonable to expect that nasality is also multidimensional.

Zraick et al. (2000) conducted a study against the null hypothesis that the perception of nasal speech quality is unidimensional. A total of 300 /i/ vowel stimuli, including one oral and four nasal vowels, were synthesized at five fundamental frequencies (80 Hz, 120 Hz, 180 Hz, 220 Hz, and 300 Hz) with a duration

of 1.5 second using a Klatt formant synthesizer (Zraick et al., 2000). Twelve clinicians participated as listeners to rate the dissimilarity of oral-nasal vowel pairs on a 7-point EAI scale, where a rating of 1 corresponded exactly same and 7 corresponded to most different. Listeners‘ judgments were analyzed using a non-metric individual differences MDS (multidimensional scaling) model, where the measure of goodness-of-fit suggested that a 3D solution accounted for approximately 82% of variance, which was about 12% more than that could be accounted for by a 2D solution. The coordinates on each dimension was calculated for each stimulus using the MDS algorithm and the correlations between these coordinates and the known acoustic parameters were examined. When certain acoustic parameters were shown to correlate with a dimension, a multiple regression analysis was used to identify the specific relationship between the dimension and the corresponding acoustic parameters. Examination of the correlation matrix suggested a fairly high (i.e. $> .70$) positive correlation between the selected acoustic variables and the corresponding dimensions. Specifically, the multiple regression analysis suggested that listeners' rating of similarity can be explained primarily by nasal voice quality (determined by the frequency of the first nasal zero (FNZ), the bandwidth of the first nasal zero (BNZ), the frequency of the first nasal pole (FNP), the bandwidth of the first nasal pole (BNP) and the amplitude of F1), loudness, and pitch. The three dimensions in together accounted for 83% of the variance of similarity ratings, where Dimension 1, 2 and 3 accounted for 54%, 18% and 11% of variance, respectively (Zraick et al., 2000).

1.4 Articulation of nasal sounds

1.4.1 Articulatory differences between oral and nasal/nasalized vowels

The variability of acoustic and perceptual characteristics of nasal sounds can be attributed to two factors, namely, anatomical variability of vocal tract and articulatory versatility. Although anatomical variability prevalently exists across different individuals, their speech shares some common features that categorize them into a finite number of phonemic categories. This fact suggests the possibility of a potential mechanism that is recruited spontaneously by speakers to accommodate individual anatomical differences by adapting their articulation.

With regard to articulation, it has been found by previous studies that the displacement of an articulator can be compensated for by adjusting other articulatory positions. For example, Hughes and Abbs (1976) examined the inter-articulator coordination of the labial-mandibular system and found the speakers could spontaneously adjust their labial displacement to compensate for the disturbed jaw movement. Additionally, Maeda (1990) found compensatory coordination between jaw position and tongue-dorsal placement for

unrounded vowels and between jaw and lip positions for rounded vowels.

The production mechanism that accommodates anatomical variability and articulatory disturbances suggests that speakers have a tendency to make spontaneous adjustments of their articulation to achieve a certain speech goal. With regard to vowel nasalization, the way of coordination between the velar gesture and the rest of the articulators and its effect on the acoustics of nasal vowels is of interest. Although the majority of previous studies assumed the same oral articulatory gestures (except velopharyngeal opening) between oral and nasal vowel counterparts, few provided direct articulatory evidence to support this assumption. On the other hand, studies on languages with phonemic distinction between oral and nasal vowels (e.g., French, Hindi, etc.) found a variety of oral-nasal articulatory distinctions. For example, tongue body is retracted in French nasal vowels /$\tilde{a}$/ and /$\tilde{ɔ}$/ (Zerling, 1984) compared to their oral vowel counterparts; French nasal vowels generally correspond to a larger front oral cavity (Engwall et al., 2006) with respect to the production of oral vowels; tongue is lower and retracted for French nasal vowel /$\tilde{ɛ}$/ versus oral vowel /ɛ/; and tongue body is generally more anterior for nasal versus oral vowels in Hindi (Shosted et al., 2010).

As nasalization is influenced by both oral articulation and velar opening, Delvaux (2003) suggested that the difference of vocal tract shape between a nasal vowel and its oral counterpart is attributed to not only lowering of the velum, but also a set of articulatory changes on tongue shape and lip opening. In order to examine the inter-speaker variability in nasal vowel articulation, Engwall et al. (2006) considered the effect of two factors, which are the size of velar opening and the total acoustic masses (determined by the volumes of the oral and nasal cavities), on nasal vowel articulation. The images of vocal tract and nasal cavity of two male and two female participants were acquired using MRI when they produced four sustained oral vowels /a, o, ɛ, œ/ and four nasal vowel counterparts. For each speaker, a series of images, including sagittal images of the oropharyngeal tract, transverse images of the velopharyngeal port, and coronal images of the nasal tract, were obtained, whereas the number of images was determined by the sampling rate and the actual size of the anatomical structure. Nasal and oral vowels were compared in terms of area function and velopharyngeal port opening quotient (VPOQ). VPOQ was defined as the area ratio between the nasal passage and the oral passage in the transverse plane where two separate air ways appeared at the velopharyngeal port. The mean velopharyngeal port opening quotient was also calculated for each vowel as the average VPOQ of all image slices with two visible passages. The definition of VPOQ facilitated not only the comparison of velopharyngeal opening degree between oral and nasal vowels but also the comparisons between different oral/nasal speech samples.

It was found that the two female subjects made a larger difference between nasal and oral vowels than the male speakers in terms of velopharyngeal opening (Engwall et al., 2006). Besides, both VPOQ and the nasal-oral difference of VPOQ were found to be largest for nasal /$\tilde{o}$/, in particular in subject F1, and the

nasal-oral VPOQ difference was smallest for the /œ̃-œ/ pair. The relatively large VPOQ for nasal /õ/ was explained as a result of the particularly narrow oral passage. Comparison of the nasal-oral difference of oral area function suggested that, in subject F1, nasal vowels were produced with more constricted pharynx and more opened oral cavity than the oral counterparts, with the only exception of the open posterior nasal vowel /ã/. In subject M2, nasal /ɛ̃/ was more constricted near the lips compared to oral /ɛ/, whereas other nasal vowels were more open in this area than their oral counterparts. In subject M1, the overall nasal-oral area difference was quite similar for all four vowels, including a constriction in the upper pharyngeal cavity and an expansion in the oral cavity (especially regions right in front of velum and near lip opening) in nasal vowels. In subject F2, the only substantial nasal-oral difference was a smaller area behind the velum for nasal /õ/. These differences of oropharyngeal articulation between oral and nasal vowel pairs were suggested as the results of changes of tongue shape and placement (Engwall et al., 2006), which change the resonant properties of the oropharyngeal cavity and, consequently, alter the acoustic characteristics of the nasal vowels.

Engwall et al. (2006) summarized four articulatory strategies used by the speakers to achieve the oral-nasal contrast found in the study: 1) Subject F1 who has the smallest nasal cavity made larger changes in both oropharyngeal tract configuration and velopharyngeal port opening than other subjects , 2) Subject F2 who has relatively larger nasal tracts achieved the nasal-oral contrast by either increasing velopharyngeal port opening and maintaining oropharyngeal articulatory placement or 3) adjusting the volume of the oral cavity with small changes at the velopharyngeal port, the pattern of which was also found in Subject M2, and 4) Subject M1 used an intermediate strategy to combine relatively small changes of velopharyngeal port opening and oral articulation. Therefore, it seems that speakers were able to spontaneously make articulatory adjustments to accommodate inter-speaker anatomical variabilities, which might be related to the highly adaptable speech motor control mechanism. The findings of Engwall et al. (2006) suggested that anatomical variability, individual articulatory strategies, and different vowel types all had influences on the acoustic characteristics of French nasal vowels.

Given the articulatory differences between phonemic nasal and oral vowels in French, it is suspicious to assume contextual nasalized vowels in English have the same articulation as oral vowels. In fact, Carignan et al. (2010) found in an EMA-aerodynamics-acoustic study that tongue body was significantly higher in the nasalized English vowel /i/ than its oral counterpart, whereas the center-of-gravity (COG) in the region around F1 for nasalized /i/ was not significantly different from the F1 COG for oral /i/. This finding suggested the possibility of a potential lingual compensatory strategy developed spontaneously by English speakers to compensate for the acoustic effect of nasalization in the F1 region. In general, the four speakers in the study of Carignan et al. (2010) tended to resist the phonologization of nasalization of /i/ through lingual compensation so that the acoustic features of nasalized /i/ would not go beyond the phonemic boundary. The

study provided evidence of languages without phonemic nasal vowels (i.e., English) against the assumption of identical articulation between oral and nasalized vowels. Furthermore, speakers were found to be able to spontaneously adjust their lingual placement to attenuate the acoustic effects of velopharyngeal opening.

Rong and Kuehn (2010) simulated vowel nasalization based on a transmission line model of the vocal tract. The lowest four formant frequencies of oral /i/ were extracted and applied as the acoustic targets for an articulatory synthesis model. By setting velopharyngeal opening to a constant moderate area, other articulatory parameters, including the positions of tongue tip and tongue body, lip opening and protrusion, jaw height and hyoid bone height, were adjusted to adapt the acoustic output to the target. In this way, even under the influence of velopharyngeal opening, the oral formant features were preserved and the nasal acoustic features were attenuated as much as possible after such articulatory adjustments. The corresponding acoustic output with/without articulatory adjustments were synthesized as the adjusted/original nasal vowels. Spectral analysis of the synthetic nasal vowels showed that, in general, the amplitude and bandwidth of low-frequency oral formants were preserved and the nasal spectral peak at around 1000Hz was attenuated in the adjusted nasal vowels, compared to the original nasal vowels. Therefore, it was suggested that oral articulation, if adjusted properly, can compensate for the acoustic effect of velopharyngeal opening (Rong and Kuehn, 2010). Moreover, as the placement of different articulators can be adjusted and coordinated in such a way that adapts the acoustic output of the model to a specific target, this approach, if generalized, might shed light on the development of a potential therapeutic approach for hypernasal speech treatment.

Carignan et al. (2010) and Rong and Kuehn (2010), using real speech and computer simulation respectively, demonstrated the acoustic effect of nasalization on the high-front vowel /i/ can be attenuated by adjusting oropharyngeal articulatory placement (especially lingual position). A relevant question is why /i/ plays such a special role among all the other vowels. House and Stevens (1956) suggested that nasalized /i/ required less velar lowering to be perceived with the same degree of nasality than nasalized /a/. In this sense, the acoustic effects of nasalization might be more salient for /i/, making articulatory adjustment more likely to happen with a goal of reducing the perception of nasality. This question can be further discussed from a perceptual view by comparing the ratings of nasality between high-front and other vowels. Watterson et al. (2007) examined the inter- and intra-listener reliability of nasality ratings for high front (HF) and low back (LB) vowels. As early acoustic studies using vocal tract analogs and acoustic analysis demonstrated the acoustic characteristics of high nasal vowels were more distorted compared to low nasal vowels due to increased oral impedance associated with a high tongue position (Andrews and Rutherford, 1973; Coleman, 1963; Fant, 1960; House and Stevens, 1956; Kent, 1966; Schwartz, 1968), the perceptual construct of nasality might also vary in high and low vowel contexts.

Two listeners (A and B) rich in experience of diagnosing and treating cleft palate speech performed

as listeners (Watterson et al., 2007). The speech samples consisted of 50 audio recordings, including 25 sentences that contained only HF vowels plus consonants and 25 sentences that contained only LB vowels plus consonants. All of the speech samples were collected from 25 children, 20 of whom were diagnosed with hypernasal speech and 5 with normal resonance. The listeners were instructed to rate the nasality scores of the stimuli using the DME method with an anchor stimulus with a moderate degree of hypernasality. The stimuli were presented to the listeners twice in different orders. To assess inter-rater reliability, the difference of equalized nasality scores between listener A and B was first calculated and then compared between HF and LB vowels. The inter-rater reliability was found to be significantly different for HF and LB vowels. To assess intra-rater reliability, the difference between the first and second ratings made by each listener was compared between the two types of vowels. A t-test showed Listener A's first and second ratings of nasality for LB vowels were more similar compared to HF, whereas listener B did not have such a difference between HF and LB vowels. In general, HF vowels were rated to be more nasal with large inter- and intra-rater reliabilities with respect to LB vowels.

Watterson et al. (2007) argued that the perceptual reliability of nasality was influenced by speech content. Specifically, the nasality ratings of HF vowels were less similar between the two listeners, suggesting that listeners are more sensitive to subtle variations of nasality in HF vowels. Therefore, it is possible that certain acoustic changes of HF nasal vowels (e.g., /i/) caused by articulatory adjustments can lead to significant changes of nasality percept.

1.4.2 Articulatory measurement and synthesis

Articulatory synthesis

Speech production and perception are correlated through the link of speech acoustics. To uncover the effect of velopharyngeal opening and oropharyngeal articulation on the percept of nasality requires a way that provides a direct control of articulation and is able to examine the effect of such articulatory control on the acoustics and perception of nasal speech. Synthetic speech meets these requirements and therefore, serves as more appropriate speech stimuli than real speech in examination of the relationship between oropharyngeal articulation and perception of nasality. This is because: 1) in natural speech, speakers do not control their articulations in a way to produce a systematic variation of a specific articulatory position in isolation unless a special training is provided (Rubin et al., 1981), resulting in a lack of direct control on individual articulators; 2) it is difficult to correlate the acoustic features of real speech to specific articulatory gestures due to the well-known one-to-multiple mapping relationship between speech acoustics and articulation; 3) a speech synthesizer allows for a partition of the articulatory configuration into a limited number of independent

control parameters (in an articulatory synthesizer) or a partition of the acoustic space into a set of formant frequencies (in a formant synthesizer) to simulate the speech production mechanism. As an articulatory synthesizer allows for direct control on a limited set of variables that control the positions of the functional articulatory units (e.g., tongue, jaw, lips, velum, etc.), it is possible to adjust the articulatory configuration by coordinating the articulatory variables to achieve a desired acoustic/auditory target.

Generally, there are five "levels" in articulatory synthesis, including 1) specifying the positions of individual articulatory variables, 2) simulating the vocal tract configuration based on the placement of individual articulators, 3) computing the cross-sectional areas of the vocal tract, 4) calculating the transfer function based on the transmission line theory, and 5) synthesizing speech based on the transfer function derived in 4) and the user-specified excitation source (Rubin et al., 1981). In addition, to synthesize continuous speech requires time-varying control over articulatory configuration and excitation source in a similar way to a key-frame animation, in which the property of each "frame" is specified along with time.

The positions of articulators can be adjusted either manually by the user or automatically by the optimization algorithm built in the synthesizer to approach the acoustic target with the articulatory model. The optimization procedure that adjusts the articulatory placement takes multiple feedbacks from different levels of the synthesizer to minimize the error between the acoustic output of the model and the target by coordinating the articulatory parameters. Based on the articulatory configuration, the vocal tract is modeled as a series of cylindrical tubes with uniform elliptical cross-sectional surfaces. The cross-sectional area of each tube is determined by the midsagittal distance (controlled by speech-dependent articulatory placement) and transverse dimension (speech-independent constant values) of the cylinder. In this way, the transfer function can be calculated based on a transmission line model of vocal tract. The non-ideal conditions, such as non-ideal termination and yielding vocal tract walls, are taken into account by sophisticating the model with special features such radiation impedance and wall vibration that simulate the relevant physical properties. Finally, a voice source is generated by specifying the intensity, fundamental frequency and some additional parameters related to the shape of glottal pulses. According to the source-filter theory (Fant, 1960), a speech sound can be synthesized with an excitation signal serving as a source and a vocal tract with a specific shape playing the role of a filter.

The primary advantages of using an articulatory synthesizer to examine the effect of oropharyngeal articulation on the perception of nasality include 1) direct control on the positions of individual articulators and 2) a straightforward relationship between articulatory placement and acoustic features provided by the simulation of a transmission line model of the vocal tract. On one hand, the coordination of articulatory positions determines the configuration of the vocal tract and herein, simulates the acoustic features of speech. On the other hand, the acoustic features of the simulated speech provide a feedback signal to lead the

articulatory adjustment in such a direction that minimizes the acoustic discrepancy between the model and the target. Another advantage is the ability of simulating continuous speech with discrete articulatory targets. Given a series of "snapshots" of the vocal tract during a speech production sequence to serve as targets, the dynamic trajectories of articulatory movements can be simulated through a physiology-based smooth interpolation between each two static articulatory targets. Therefore, the design and implementation of an articulatory synthesizer provides a convenient and interactive tool for examining the relationship between speech articulation and acoustics.

Articulatory measurement

Ideally, to recover the articulatory motion during running speech requires dynamic imaging that has sufficient spatial resolution to cover the entire vocal tract in 3D and a fast acquisition rate to capture the speech kinematic movements. The ideal acquisition rate should be at least 100 frames per seconds for continuous speech and greater than 40 Hz for syllables. In addition, to explore the articulatory space of a speaker requires recording of a large speech corpora from different speakers, as per Aron et al. (2006). With consideration to all these issues, the acquisition technique needs to be fast, accessible and inexpensive at the same time. No single imaging technique would meet all these criteria so far (Engwall, 2000a). MRI offers a good spatial resolution of the vocal tract but is relatively slow in acquisition, preventing it from capturing fast speech movement. Ultrasound provides a higher acquisition rate for tracking of tongue body movement but is unable to track the apex of the tongue. EMA is used for recording and real-time display of articulatory movement but it only provides sparse information on articulatory configuration (Aron et al., 2006). Therefore, to construct an articulatory model, recent studies have been combining two or more techniques above to acquire the speech corpora.

Aron et al. (2006) coupled electromagnetic (EM) sensors and ultrasound (US) for tongue tracking. US images were used to acquire the tongue contour and EM data completed the acquisition of tongue shape by measuring the apex region. The curve that best matches the acquired tongue contour and passes through the EM sensor location was regarded as the 2D recovery of the tongue shape by using interpolation and regulation schemes. Aron et al. (2006) suggested the setup for the combination of EM sensors and US images was reproducible. In this way, the combination of EM and US can be used to recover tongue grooving above the hyoid bone in dynamic speech.

Engwall et al. have been working on combination of different imaging techniques in speech modeling and articulatory analysis. In Engwall (2000a), the influence of artificial sustaining of articulation during static MRI acquisition was assessed by comparing the static articulation acquired by MRI with the dynamic measurements from an EMA-EPG study. It was found, rather than showing non-representative articulations,

the MRI-based articulatory measurements represented a case of hyper-articulated speech. Regardless of substantial differences concerning jaw position, lip protrusion and tongue configuration between the two studies, static MRI still provides valuable information and reference structures for dynamic speech. The major divergence between static and dynamic articulations was related to lip protrusion and jaw height and the tongue position became more neutral when it was constrained to a static position. These divergences were explained in Engwall (2000a) as the results of hyper-articulation in static MRI acquisition. Therefore, it is possible to use static MR images as the target for model construction, but the hyper-/hypo-articulation control theory is needed to determine to what extent the target can be reached while simulating a running-speech-like output with the articulatory model.

Although the capability of representing dynamic speech movement has been demonstrated in Engwall (2000a), static MRI data still needs to be complemented with dynamic speech measurements to capture continuous speech movement and, furthermore, generate a model representative of running speech. Engwall (2001) used EMA data collected with the Movetrack measurement system to provide dynamic articulatory movements in a three-dimensional tongue model developed based on the MRI images of a subject producing 43 artificially-sustained articulations of Swedish. The EMA measurements were first scaled to fit the tongue model and the corresponding control sequences were then generated from the EMA measurements for each articulatory parameter of the model. This results in a three-dimensional tongue model, whose shape and articulatory parameters are determined through statistical analysis on the MRI data, whose activation pattern is derived based on the combination of MRI and EMA data, and whose dynamic movements are determined by the EMA measurements.

Engwall (2000b) combined EMA and EPG to examine the dynamical aspects of coarticulation in five Swedish fricatives in different vowel contexts. Simultaneous EMA and EPG recordings enable measurements of jaw motion, lip protrusion and tongue movements with EMA and the linguopalatal contact with EPG. EMA and EPG were demonstrated to be compatible without prominent interactions that cause large measurement errors (Rouco and Recasens, 1996). Therefore, the combination of EMA and EPG was recommended as one of the standard measurements of dynamic speech articulation (Hardcastle et al., 1996; Hoole et al., 1993; Perkell et al., 1999). EMA, which measures three dimensional point-wise articulatory positions, is a preferable technique in acquiring dynamic articulatory information compared to other instruments such as X-ray. The ethical constraints linked to the radiation of X-rays prevent it from extensive research applications, regardless of its fast sampling rate and good spatial resolution. EPG provides dynamic and point-wise binary information on tongue-palate contact. The main drawback of EPG is related to its low spatial resolution, that is, only binary information on tongue-palatal contact is provided. Nevertheless, its compatibility with other imaging systems such as ultrasound (Stone et al., 1992; Stone and Lundberg, 1996), MRI (Alwan

et al., 1997; Narayanan et al., 1997) and articulography (Nguyen et al., 1998), largely extends its application as a supplementary tool. Engwall (2000b) used combined articulography and EPG to measure dynamic and point-wise articulatory information to examine the temporal aspects of speech production and coarticulation. The measurements in Engwall (2000b) were also used for assessment of the validity of using static MRI to represent dynamic speech features in Engwall (2000a). Therefore, if the three techniques (i.e., MRI, EMA, EPG) can be combined to provide articulatory information that is both continuous and dynamic, it will be an ideal tool for articulatory modeling. However, the current technology does not allow for combination of EMA and MRI, or EPG and MRI due to the artifact that would be expected from the EMA sensors and the EPG electrodes within the MRI scanner.

Moen and Simonsen (2007) used combined EMA and EPG to study the Norwegian coronal stops. In addition to linguopalatal contact, EPG frames were also used to determine the onset and offset of the consonantal closure, which can assist segmenting and interpreting EMA data. Moreover, EPG informs the location of tongue tip when it contacts the palate to produce a coronal stop closure. Such information on the tongue tip position is not available from the EMA measurements because the most anterior sensor on the tongue is usually placed a certain distance (e.g. $0.5 - 1cm$) behind apex to prevent interference with speech. EPG provides lingual articulatory information when contacts happen and herein, supplements the EMA articulatory measurements. In addition, the shape of the EPG palate provides a direct recovery of the hard palate of the speaker. Moen and Simonsen (2007) claimed the combined use of EMA and EPG gives a clearer and more comprehensive view of what is going on in terms of articulation inside the mouth during continuous speech.

Vocal tract modeling and speech synthesis based on articulatory measurements

According to the previous discussion, articulatory synthesis relies on the adjustment of a set of independent articulatory parameters that control the movement of the primary articulators. The articulatory parameters are partitioned in such a way that accounts for as much variance of articulatory movement as possible. Built on different mechanisms, there are two types of articulatory models.

The first type uses a statistical model based on a large inventory of vocal tract configurations that cover the articulatory space of a speaker involving all possible articulatory gestures in different speech sounds. The vocal tract configuration can be represented in numeric forms such as area functions and mid-sagittal distances. Factorial analysis, such as the Principal Component Analysis (PCA), is used to derive the primary components that, taken together, account for the majority of variance of speech articulation in the inventory. Each eigenvalue in PCA corresponds to the amount of variance accounted for by a factor, which is derived by PCA as the eigenvector corresponding to the eigenvalue.

As the goal of PCA is to extract a minimum number of factors that account for the majority of variance in the data, there is no guarantee that the derived components (factors) are interpretable in articulatory terms. In order to derive the articulatory components of the vocal tract, an assumption is made such that the complex activities of the articulators can be represented by a limited set of independent and functional blocks, which determines the shape of the vocal tract. Maeda (1990) used more than 1000 digitized tracings of vocal tract shapes from the simultaneous radiofilm and labiofilm recordings of 10 French sentences produced by two female speakers to extract the principal articulatory components based on the arbitrary factor analysis (Overall, 1962). The jaw component was first extracted based on the measured jaw position and then subtracted from the data. The rest of the components were extracted orderly as principal components. According to Maeda (1990), it was found that PCA is not able to extract the component that accounts for the jaw motion, so the placement of jaw can not be recovered with the PCA components. Therefore, an arbitrary factor analysis was used to first extract jaw component, and then the rest of the components were derived using PCA. It was found that the components extracted by PCA all had articulatory interpretations, such as tongue dorsal position, tongue dorsal shape and tongue tip position. Similar to the extraction of components in the principal vocal tract, lip and laryngeal sections were also analyzed using the arbitrary factor analysis. The extracted components along the entire vocal tract compose an articulatory model that is controlled by seven articulatory parameters, which are jaw height, tongue placement (i.e., dorsal position and shape, tip position), lip placement (i.e., opening and protrusion) and larynx height.

Zheng et al. (2003) used the three-mode PARAFAC (parallel factors) analysis to examine the 3D tongue shape during the production of nine vowels based on the 3D MRI images of five speakers. In the analysis, both 3D and 2D measurement vectors were obtained from the coronal images of the tongue. PARAFAC analysis of both 2D and 3D coordinate data resulted in a stable two-factor solution that explained over 70% of the variance. The two factors represented the raising-lowering and anterior-posterior movements of the entire tongue in both 2D and 3D analyses. Therefore, the two-factor PARAFAC model was capable of extracting linguistically meaningful and statistically valid 2D and 3D factors from the MRI-derived data of tongue shape. The interaction between speaker identity and vowel identity was able to be modeled with the two factors of PARAFAC. The resulting vowel space was demonstrated to be consistent with the traditional height-fronting relationship of tongue.

Another type of articulatory model is biomechanics-oriented. The design of this kind of model is determined by 1) the positions of the primary articulators, including tongue tip, tongue body, tongue back, lips, jaw, hyoid bone, etc., and 2) the biomechanical relationship between these articulators. Different from the previous kind of model, this type of articulatory model does not require a full image of the vocal tract but, instead, is established based on the position of a few freshpoints on the primary articulators. The vocal tract

shape can be estimated from the positions of the freshpoints based on the biomechanical relationship among the articulators. As point-wise articulatory data is easy to acquire with EMA, it is possible to record a large database with relative low cost and no health hazard.

The second type of model is commonly used in articulatory synthesizers. However, a standard vocal tract model with a fixed anatomical framework is not completely consistent with a real vocal tract in that it cannot account for the large inter-speaker variability of anatomical structures and articulatory dimensions. Such inter-speaker variability might increase the error of articulatory recovery during inverse filtering due to the discrepancy between the model framework and a speaker-dependent vocal tract. Therefore, a normalization procedure is needed to adapt the "standard" articulatory model to different speakers by taking into account inter-speaker anatomical variability (e.g., size and shape of anatomical structures) and articulatory dimensions.

McGowan and Cushing (1999) proposed a normalization procedure for articulatory recovery using analysis-by-synthesis. First of all, Normalizing Map (NM) was defined as the transformation of the midsagittal shape or articulatory positions of a speaker-dependent vocal tact to a "standard" vocal tract model. According to the definition of NM, the goal of normalization as described above is actually an inverse NM. In addition to the inverse NM, another type of normalization should be conducted to extract the parameters of the standard vocal tract model that are not affected by the (inverse) NM and then adjust these parameters to fit the corresponding dimensions of individual speakers. These parameters account for the anatomical features that are not accounted for in the space of NM such as the length and transverse dimensions of the vocal tract. McGowan and Cushing (1999) defined such parameters as PONM (Parameters Orthogonal to the space of NM). During the second step of normalization, the PONM are adjusted so that the configurations of the vocal tract model and a speaker-dependent vocal tract mapped to one another under the NM or its inverse produce the same acoustic outputs (Badin et al., 1995; Beautemps and P. Badin, 1995). McGowan and Cushing (1999) used the ASY model of the Haskins Laboratories articulatory synthesizer as a standard vocal tract for normalization. The X-ray microbeam speech production database was used to serve as the articulatory database for model adaptation. To extract the PONM, a set of vowel samples were selected from the database to serve as training stimuli. The midsagittal coordinates of the articulators in the ASY model were adjusted so that a close visual match was achieved between the midsagittal contour of the ASY model and the pellet positions of the articulators for the stimuli selected from the database. The orientation of the pellet coordinate system was then adjusted to match the ASY model coordinate system. Minor adjustments were made to the raw positions of some pellets to make the closest match with the corresponding ASY landmarks. After the manual adjustments, a circle with a 2cm-radius was fitted to the posterior part of the tongue so that it passes through the three rearmost pellets on the tongue surface, where the center of the

circle was marked to serve as the center of the ASY tongue model. The tongue model was further completed by drawing a line from the frontmost pellet on the tongue surface to a point on the posterior tongue circle, the location of which is determined by the angle between the line and the jaw vector, which should has a constant value of 0.55π. The hyoid coordinates of the ASY model were left with the default values and the nasal port was set to be closed.

The PONM to be adjusted include the length of the vocal tract (i.e., vtln) and the transverse dimension of the pharynx, both of which should be speech- and speaker-dependent anatomical features. As the cross-sections in the pharyngeal region of the ASY model are represented as ellipses with the dimension of one (i.e., midsagittal) axis equal to the midsagittal distance and the dimension of the other (i.e., transverse) axis set to be fixed, the inter-speaker variability of the transverse axis dimension along the tract should be accounted for by adjusting the relevant PONM. Two parameters "trans_phar_inter" and "trans_phar_slope," standing for the intercept and slope of the line fitted to simulate the relationship among the transverse dimensions of three sections of of the pharynx, served as PONM. In the soft and hard palate regions of the ASY model, the cross-sections were modeled as ellipses with the length of the transverse axis as a function of the midsagittal distance: transverse length = COEFF$*$(midsagittal distance)$^{\mathrm{POW}}$, where COEFF and POW are regarded as the corresponding PONM, where in soft and hard palate regions, these two parameters were soft_pal_coeff, soft_pal_pow and hard_pal_coeff, hard_pal_pow, respectively (McGowan and Cushing, 1999). The simple genetic algorithm was used to optimize the PONM with a goal of minimizing the discrepancy of the lowest three formant frequencies between the model simulation and the stimulus in the database.

Given the optimized PONM, the inverse NM was applied to the training vowel samples using artificial neural networks. The ASY articulator coordinates were rescaled to serve as the input of the neural networks and the Cartesian coordinates of the adjusted pellet positions served as the output of the neural networks. By selecting an appropriate number of neurons in the hidden layer of the networks and a proper error criterion, the PONM were adjusted. Finally, McGowan and Cushing (1999) performed a task-dynamic recovery procedure to acquire the task-dynamic parameters, based on which the dynamic articulatory positions in the database were recovered in the ASY model with the adjusted PONM and the inverse NM acquired in the training. The goodness-of-fit of the PONM was evaluated by comparing the recovered articulatory trajectories with the original pellet positions in the database. McGowan and Cushing (1999) concluded that the adjustment of PONM allowed for better simulation of speaker-dependent articulation. Given the adjusted PONM, it is possible to construct an inverse NM for each individual with satisfactory accuracy of articulatory fitting.

Naito et al. (2003) proposed a speaker normalization method based on a speech generation model that can achieve high-performance speaker adaptation with relatively small amount of data. As the shape of

the vocal tract represented by area functions is both speaker- and phoneme-dependent, the normalization is first conducted phoneme-wise on a standard articulatory model. Phoneme-dependent area functions were estimated by adjusting the articulatory parameters of the model in such a way that the shape of the vocal tract model matched the measured area function of the vowel. The limits of the oral and pharyngeal sections of the vocal tract in the standard phoneme-dependent model were then independently stretched to approximate the phoneme-dependent area functions of a target speaker. The relationship between the lengths of the oral and pharyngeal cavities (i.e., l_1, l_2) and the lowest three formant frequencies (i.e., F1, F2, F3) was drawn as a map. In the map, the speaker-dependent factor l_1 and l_2 of the vocal tract that are directly related with F1, F2, F3 could be estimated based on measurements of formant frequencies. After the lengths (l_1, l_2) of the target speaker's vocal tract were normalized, a frequency warp function was computed for each phoneme in such a manner that correlated the formant frequencies (F1-F7) simulated with the model of the normalized target speaker with the corresponding formant frequencies simulated with the standard vocal tract model. The seven formant frequencies were linearly interpolated to achieve a continuous function. As a result, the recognition error of the synthetic speech after the application of the proposed normalization was reduced by about 13%.

1.4.3 Velopharyngeal aperture estimation

Apart from the oral articulatory movements (e.g., tongue displacement, jaw raising/lowering, lip opening/closing and protrusion/retraction, etc.), which can be measured by various imaging approaches as discussed above, velar movement is the critical articulatory event related to nasalization. When the velum is raised to form an air-tight seal at the velopharynx, nasal cavity is shut from the oropharyngeal tract, so the air from the pharyngeal cavity only flows into the oral cavity. On the other hand, when the velum is lowered, nasal and oropharyngeal tracts are coupled at the velopharyngeal port to allow airflow through both oral and nasal cavities. If the velopharyngeal mechanism is impaired in such a way that the valve cannot fully close when it is supposed to, a condition known as "velopharyngeal incompetency (VPI)" can develop. Velopharyngeal incompetency can be caused by anatomical abnormalities such as cleft palate and neuromuscular dysfunction related to the velopharyngeal (VP) mechanism. Therefore, estimating the size of velopharyngeal aperture is critical in assessing the velopharyngeal function of individuals suspected to have velopharyngeal dysfunction. However, it is more difficult to measure VP opening area compared to other articulatory gestures due to the complex biomechanics of the velopharyngeal structures. In this section, the currently-available approaches of estimating velopharyngeal opening (VPO) area are discussed.

Generally speaking, velopharyngeal opening area can be estimated in two ways: 1) using imaging tech-

niques such as MRI to directly image the velopharynx and estimating/measuring the size of VP opening area based on the image; 2) using the pressure-flow technique (Warren, 1964) to estimate VPO based on the aerodynamic measures (i.e., transvelic pressure and nasal airflow).

Imaging techniques enable visualization of the velopharynx so that direct measurement can be made on the VP area. Among the commonly-used imaging techniques, X-ray exposes radiation to human and is limited for research applications. Regardless of the potential health hazard, X-ray only provides 2D visualization of the vocal tract, so assumptions need to be made on transverse dimension in order to estimate the VP area. On the other hand, MRI has been used commonly in previous studies for vocal tract imaging due to its safety and capability of imaging in three dimensions. Story (1995); Story et al. (1996) used 3D MRI to derive the area functions of eleven vowels, one liquid, three nasal consonants and three stops produced by a subject. The nasal coupling areas were estimated based on the images at the velopharynx, which were around $1.04cm^2$, $1.26cm^2$ and $1.09cm^2$ for /m/, /n/ and /ŋ/, respectively. However, as discussed above, the sampling rate of 3D MRI is not sufficient to capture details of dynamic speech movement, so the acquisition of 3D images in Story (1995); Story et al. (1996) was enabled by artificially sustaining the articulation. Artificial sustaining of articulation during MRI imaging is suggested to lead to hyper-articulation (Engwall, 2000a), which might result in a discrepancy of the VP area compared to a natural speech condition. In other words, the trade-off between spatial and temporal resolution of 3D MRI prevents it from capturing the time-varying velar movement in dynamic speech. 2D MRI, on the other hand, has a faster acquisition rate, but encounters the same problem as X-ray, that is, only midsagittal distances rather than area functions are available from the images. Articulography has a fairly high sampling rate (e.g., 200Hz) sufficient to capture dynamic speech movement in three dimensions, but only sparse articulatory information represented by the motions of a few freshpoints on the primary articulators can be measured, so the velopharyngeal aperture area is not directly available from the articulography measurement either.

The pressure-flow technique provides an alternative way to estimate VP area based on aerodynamic measures using the hydrokinetic method (Warren and Dubois, 1964). The aerodynamic measures include intraoral and nasal pressures and nasal airflow rates.

Warren and Dubois (1964) designed an experiment to measure the speech aerodynamics (i.e., oral-nasal pressure difference and nasal air flow) required by the hydrokinetic method. The oropharyngeal pressure was measured by a small balloon, which passed through the left nostril and the VP orifice of the subject into the oropharynx. The nasal airflow was measured by a heated pneumotachograph, which was connected to the right nostril through a large snugly-fitting tube. In order to obtain the differential pressure at the VP orifice, nasal pressure was estimated based on nasal airflow and nasal resistance, where the nasal resistance was measured during expiration and assumed to stay constant during speech. The estimated nasal pressure

was subtracted from the oropharyngeal pressure to serve as the orifice differential pressure. Given the nasal airflow rate and orifice differential pressure, VP orifice area can be estimated based on the hydrokinetic equation ($A = \frac{\dot{V}}{0.65\sqrt{2(\frac{P_o - P_n}{D})}}$, where A is the estimated VP orifice area, $\dot{V}$ is nasal airflow rate, P_o and P_n are oropharyngeal and nasal pressures, and D is a constant representing air density.)

Throughout the recorded signal, measurements were only made at the points where peak pressures occurred during the consonants and at the midpoint of the voiced interval of the vowel (Warren and Dubois, 1964). Velopharyngeal opening (VPO) areas were estimated only at these selected time frames. Specific time frames were selected to guarantee the success of applying the hydrokinetic method as the hydrokinetic equation might not work for every time frame during continuous speech. For example, when VP orifice was open too wide, the nasal pressure might be equal to or even larger than the oropharyngeal pressure, in which case the differential pressure goes negative, resulting in a failure of the hydrokinetic equation. Even within the selected time frames, Warren and Dubois (1964) still found certain cases where the pressure component due to nasal resistance was equal to the recorded oropharyngeal pressure. When such a situation occurred, an arbitrary value equal to the lowest pressure that is measurable at the highest amplifier sensitivity was assigned as the differential pressure. Therefore, in such cases, the actual VP orifice area was larger than the estimated value due to the over-estimated differential pressure. Another issue of VPO estimation during continuous speech is related to the signal produced by the pressure transducer, which is composed of aerodynamic events related to both functional velar movements and oscillations introduced by small non-functional velar movements. Functional velar movements contribute to the speech production mechanism, whereas non-functional movements are trivial in a sense that they do are not directly related to the outcome of speech. Andreassen et al. (1991) found small VP openings in oral sounds based on the hydrokinetic estimation, which most likely reflect non-speech-related velar movements in the presence of an air-tight velopharyngeal closure. Such small velar movements result in nasal airflows that would estimate small orifice areas, whereas functional openings may not actually occur. To eliminate the effect of such non-functional velar movements on the estimated VPO, the aerodynamic signals should be low-pass filtered to retain only the envelope of the signal, which reflects the overall airflow/pressure variation related to speech production.

Andreassen et al. (1991) measured the oral-nasal differential pressure and nasal airflow rate at the peak pressures in CV syllables /pi/, /pa/, /fi/ and /p/ in "hamper" produced by 20 speakers with normal speech. The peak nasal airflows and oral-nasal differential pressures at the nasal airflow peaks were measured in CV syllables /mi/, /ma/ and /m/ in "hamper." The volume rate of airflow through the nose was measured by a pneumotachometer coupled to the least resistive nostril of the subject via a tube inserted into the anterior nasal cavity. The pressure drop across the velopharyngeal orifice was measured by two catheters, one coupled to the other nostril of the subject and the other inserted into the subject's mouse. Andreassen

et al. (1991) estimated the VP areas at the selected time frames of these tokens. The typical VP areas were found to be no more than $1mm^2$ in /pi/ and /fi/ and no more than $5mm^2$ and $3mm^2$ in /pa/ and /p/ in 'hamper', respectively. For nasal tokens, the typical VP areas were no less than $20mm^2$, $27mm^2$ and $13mm^2$ for /mi/, /ma/ and /m/ in 'hamper', respectively. Comparisons between the VP areas estimated by the pressure-flow method in Andreassen et al. (1991) and the direct measurements based on MRI (Story, 1995) suggest that the pressure-flow technique tends to under-estimate the VP area compared to the measurement based on static 3D MRI. Such discrepancies might be attributed to potential hyper-articulation caused by artificial sustaining of speech during static MR imaging. Significant aerodynamic differences (i.e., oral-nasal differential pressure/nasal airflow rate/estimated VPO) were found both between genders and between vowel contexts, suggesting a necessity to examine individual patterns.

Warren and Dubois (1964) and Andreassen et al. (1991) both used the pressure-flow technique to estimate VP orifice area at selected time frames during continuous speech or repetitive syllables. The pressure-flow technique has been demonstrated to be an effective approach to estimate VP opening area alternative to direct measurements provided by imaging techniques. However, instrumentation could be a potential issue of applying the pressure-flow technique. Inserting a balloon through the nostril as had been done in Warren and Dubois (1964) could cause considerable physical discomfort of the subject. In another study by Warren (1967), an improvement on instrumentation similar to what was used in Andreassen et al. (1991) was applied. However, as very few technical details were provided on the instrumentation regarding where and how the catheter was placed inside the mouth in both studies, it is unknown whether the pressure captured by the oral catheter represents the desired intraoral pressure near the velum. Therefore, a more sophisticated instrumentation technique is needed to apply the pressure-flow technique.

1.5 Purpose of the study

The speech of individuals with velopharyngeal incompetency is characterized by hypernasality, which is the perceptual quality perceived by a listener as a result of excessive transmission of acoustic energy through the nose, as caused by the failure of a complete velopharyngeal closure. Anatomical defects such as cleft palates and neuromuscular dysfunctions related to the velopharyngeal mechanism are two primary causes of velopharyngeal incompetency. Although surgeries can repair the anatomical defects related to VPI, post-surgical speech treatment is often necessary to correct the inappropriate articulatory behaviors developed spontaneously by patients with VPI. Traditional therapies for hypernasal speech are in general, empirical, relying primarily on clinicians' subjective perceptual judgement and the theoretical rules to reduce nasal emission. For example, one simple approach is to ask the speakers with VPI to increase lip opening so

that the enlarged volume of the oral cavity reduces the resistance of airflow transmission inside the oral cavity and in turn, admits more acoustic energy from the subglottal system. As a result, the proportion of energy transmitted inside the nasal cavity is reduced, which should result in a reduction of the perceived hypernasality. However, large individual variabilities in terms of anatomy, articulation and acoustics make it difficult for such simplified therapies to meet individual needs. An individual-based behavioral speech therapy is needed to improve the intelligibility of hypernasal speech in different speakers by taking into account individual variability.

As the production mechanisms of hypernasal speech and phonetically-nasalized speech are similar, this study used phonetically-nasalized vowels recorded from two American English speakers to simulate hypernasal speech. This enables large-scale articulatory data acquisition by avoiding the technical difficulties related to physiological data collection within pathological populations.

The one-to-multiple mapping relationship between acoustics and articulation makes it possible to produce the same acoustic outcome with different articulatory configurations. This is enabled by inter-articulator coordination, as operated in the form of motor equivalence (Hughes and Abbs, 1976). According to Hughes and Abbs (1976)'s experiment on the coordination among upper lip, lower lip and jaw by examining the relative contribution of each of these articulators to the superior-inferior distance between the upper and lower lips, evidence of motor equivalence was found in the speech movement coordination of the labial-mandibular system. Maeda (1990) found inter-articulator compensation between the jaw position and the tongue-dorsal position in unrounded vowels and between jaw and lip positions in rounded vowels. A more general claim about motor equivalence was made by Maeda (1990) that the vocal tract configuration for the production of a consonant or a vowel can be realized through many different combinations of articulatory gestures. Presumably, the deviation caused by the shifted position of one articulator could be compensated for by repositioning articulators so that an appropriate "goal" can be achieved. Such compensatory coordination between different articulators suggests speech articulation is essentially a "goal oriented" process operated by "motor equivalence" (Gay et al., 1981). Given an acoustic/auditory target, a deviation from the target due to the displacement of an articulator can be compensated for by adjusting the placement of other articulators.

In the light of motor equivalence, the relationship between VPO and oropharyngeal articulation can be described as a question: what oropharyngeal articulatory adjustments can be made to compensate for the acoustic deviation caused by excessive VPO related to velopharyngeal incompetency?

This study constructed an articulatory model to develop oropharyngeal articulatory adjustments under the influence of VPO to compensate for the acoustic discrepancies between hypernasal and normal speech. Given an acoustic/auditory goal and the physiological mechanism of speech production, the articulatory model can simulate the "goal-oriented" process of "motor equivalence" (Gay et al., 1981). Specifically, the

"goal" can be determined either in an acoustic view to attenuate the nasal acoustic features and restore the oral formant structures, or in an auditory view to reduce the perception of hypernasality. The physiological mechanism of speech production takes into account the biomechanical and kinematic properties of the articulators as well as inter-articulator coordination.

Therefore, the first research hypothesis of this study is the acoustic deviance caused by excessive velopharyngeal opening in hypernasal speech can be compensated for by repositioning and coordinating other articulators (except the velum) to achieve the goal of attenuating the acoustic outcome caused by VPO and in turn, reducing hypernasality.

As a standard articulatory model with fixed anatomical structures and articulatory dimensions is not able to account for inter-speaker anatomical and articulatory variabilities, the articulatory adjustment developed by such a model might not work as effectively as from one to another speaker. Actually, the inter-speaker anatomical difference could be a potential factor that reduces the effectiveness of the articulatory adjustment developed by the model because part of the compensation effect in this case is devoted to reduce the acoustic deviation caused by the anatomical discrepancy between a speaker and the standard model. Such articulatory adjustment is not oriented from the goal of compensating for the shifted position of an articulator, but is made to minimize the acoustic discrepancy caused by anatomical variability. Therefore, the anatomical dimensions of the articulatory model need to be adapted to individual speakers before the development of articulatory adjustment to compensate for VPO. This will be achieved by 1) recording a speech database with both acoustic and articulatory data of a speaker producing a variety of speech sounds; 2) fitting the articulatory data into a standard model framework; 3) customizing the model with speaker-dependent anatomical dimensions by training it with the speech samples in the database to minimize the acoustic discrepancy between the model simulation and the acoustic signal; and 4) adapting the movement ranges of the articulators in the model to the articulatory spaces of individual speakers.

As the articulatory model after the speaker adaptation as described above can account for inter-speaker anatomical and articulatory variabilities, it is able to simulate the speech articulation of different individuals more realistically and, more importantly, it can provide individual-based articulatory strategies that are intended to help individuals with VPI to reduce hypernasality and improve their speech intelligibility. Therefore, the second research hypothesis of this study is that speaker-dependent oropharyngeal articulatory adjustments can be simulated to compensate for the effect of VPO with a speaker-based articulatory model. The long-term goal of this study is to apply such a speaker-adaptive articulatory model as a therapeutic tool to diagnose and correct articulatory disturbances related to VPI in clinical speech treatment.

Chapter 2

Method

2.1 Pilot study

A modeling study of the effect of articulatory adjustment on the acoustics and perception of nasalized vowels in American English

A simulation study was conducted to examine the effect of articulatory adjustment on compensating for the acoustic and perceptual outcome caused by velopharyngeal incompetency. An articulatory synthesis function (*Speech Processing and Synthesis Toolboxes*, Childers (2000)) was used to derive such articulatory adjustment by coordinating the parameters of the articulatory model, aiming at achieving the acoustic target. Given an acoustic target, three types of /i/ vowels were synthesized with different articulatory strategies, which corresponded to oral vowels, nasal vowels, and nasal vowels with adjusted oropharyngeal articulation that compensated for the acoustic deviation of velopharyngeal opening. The synthetic vowel stimuli were subject to acoustic analysis, focusing on examining the low-frequency spectral differences among the three types of vowels. A perceptual experiment was then conducted to rate the nasality of the synthetic vowels by eight listeners using a direct magnitude estimation (DME) method. The rated nasality scores were compared across the three types of vowels to examine the effect of adjusted articulation on the perceptual ratings of nasality.

2.1.1 Procedures

Participants

Three males and three females participated as speakers, all of whom speak American English as their first language. The criteria of speaker selection include capability of performing the tasks required of them and negative history of speech, language, voice and resonance problems. All of the speakers were absent of any

sign of cold or upper respiratory infection at the time of participation and their speech was evaluated to be in normal resonance conditions by the experimenter. Eight students who speak American English as their first language served as listeners. Listeners are free from any history of hearing, speech, voice, or language difficulties and were in good health condition at the time of participation.

Speech recording

The speech samples consisted of 72 words in the form of CVC (consonant-vowel- consonant) or CV (consonant-vowel), where V stands for the English vowel /i/, which was chosen due to its unique position in the vowel space as a high-front vowel with relatively large spectral distance between F1 and F2. Because the duration of the stimulus can influence the perception of nasality, the speakers were asked to sustain each stimulus for one second, which was controlled by a timed light signal. Each vowel was prolonged to some extent compared to conversational speech, but was regarded to maintain natural quality. All speech samples were recorded in randomized orders through a unidirectional microphone set and saved as wave files.

Speech synthesis

Speech stimuli for the perception task were synthesized using the articulatory synthesis function in Speech Processing and Synthesis Toolboxes (Childers, 2000). Selection of isolated vowels as stimuli was intended to exclude confounding factors such as context and coarticulation that may potentially influence the perception of nasality.

Synthesizing speech using Speech Processing and Synthesis Toolboxes includes four steps: 1) computing an acoustic target using the analysis function, 2) coordinating and adjusting the articulatory parameters to minimize acoustic discrepancy between the model simulation and the target, using the articulatory optimization function, 3) creating a voice source and 4) synthesizing speech using the synthesis function. Specifically, given a speech sample as the input, the lowest four formant-frequency (i.e., F1, F2, F2, F4) contours were computed to serve as the acoustic target. The optimization of articulation is enabled by parceling out the articulatory configuration into seven controllable articulatory parameters to represent the placement of individual articulators. The seven articulatory parameters, in control of the positions of tongue tip and tongue body, lip opening and protrusion, jaw height, hyoid bone height, and velar position, are regarded to be independent of each other. Therefore, the articulatory parameters can be adjusted individually to adapt the acoustic features of the simulated model output to the target. The goodness of fit is determined by an evaluation function, computed as the weighted sum of errors of the four formant frequencies to be minimized during the optimization. By minimizing the acoustic discrepancy between the model and the target, a set of optimal articulatory parameters was computed. An excitation source was then generated with an appropri-

ate glottal waveform and a user-specified fundamental frequency to serve as the voice source for articulatory synthesis.

In the pilot study, 72 speech samples (12 words x 6 speakers) were recorded. The vowel portion of each word was segmented, within which the fundamental frequency (f0) contour was tracked. Three vowels with least variation of f0 were selected from the 12 vowels produced by one speaker so that a total of 18 (3 selected samples x 6 speakers) vowel samples were selected to serve as the input of the analysis function to derive the acoustic targets. To eliminate the inter-speaker differences of fundamental frequency and the corresponding effect on the rating of nasality, formant frequencies were normalized with respect to the fundamental frequency for all speakers. This was done by a two-step normalization: 1) the logarithm of "relative formant frequencies" were calculated as $\log(\frac{F1}{f0})$, $\log(\frac{F2}{f0})$, $\log(\frac{F3}{f0})$, $\log(\frac{F4}{f0})$ and 2) the "normalized formant frequencies" were computed as $\exp(\log(\frac{F1}{f0}) + \log(fn))$, $\exp(\log(\frac{F2}{f0}) + \log(fn))$, $\exp(\log(\frac{F3}{f0}) + \log(fn))$ and $\exp(\log(\frac{F4}{f0}) + \log(fn))$, where fn is the "standard" fundamental frequency set to 110Hz. The normalized formant frequencies ($\tilde{\text{F}}1$, $\tilde{\text{F}}2$, $\tilde{\text{F}}3$, $\tilde{\text{F}}4$) were applied as the acoustic target for articulatory adjustment. As hypernasal speech is characterized by excessive velopharyngeal apertures, a velopharyngeal opening with a $220mm^2$ orifice was used in the model to simulate the cause of hypernasality. The VPO area was determined by a trained listener (not included in the perceptual task) to represent moderate hypernasality. There is a large leap to simulate hypernasal speech with fixed VPO in an articulatory model, but the synthesized nasal vowels were judged by the trained listener to reflect the quality of hypernasality in an acceptable way.

Three types of /i/ vowels, namely, oral vowels (O), nasal vowels (N) and nasal vowels with articulatory adjustment (NA), were synthesized. For O and NA, the articulatory parameters were adjusted in the same way except the setting of VPO, which was set to zero for O and $220mm^2$ for NA before initiating articulatory optimization. On the other hand, the VPO for N remained closed (i.e., VPO=0) during the articulatory optimization and it was not changed to open (i.e., VPO=$220mm^2$) until after finishing the optimization. Such a difference of manipulation on the articulatory parameters distinguished nasal vowels with articulatory adjustment from those without articulatory adjustment, which ensured the articulatory adjustment was made by taking into account the overall spectral characteristics in the low-frequency region (i.e., F1–F4) rather than an individual formant frequency to compensate for the nasal acoustic features. The simulation of hypernasal vowels (N) assumes no incorrect articulatory pattern developed spontaneously by the speaker, whereas nasal vowels with adjustment (NA) simulate hypernasal vowels after the oropharyngeal articulatory pattern has been adjusted towards a desired direction.

Speech perception

During the preparation of the listening task, a training session was provided in which listeners were presented with a series of synthetic vowels, not drawn from the stimuli but generated with the same synthesizer, so that listeners could familiarize themselves with the synthetic speech quality and the DME procedures. Five synthetic vowels with different degrees of VPO were presented to the listeners for practice of nasality rating. The perceptual task began after the listeners became sufficiently familiar with both the synthetic speech quality and the DME method. The majority of listeners (five out of eight) chose to proceed to the experiment after practicing once with the five training stimuli, two listeners proceeded after two runs of practice, and one listener practiced three times.

In the listening task, 54 synthetic vowel stimuli were presented to the listeners in six groups. Each group included three O/N/NA sets, where the three vowels (O, N, NA) within a set were synthesized based on the same acoustic target. Within each group, the orders of the nine vowel stimuli were randomized. The listeners were asked to rate the nasality of the stimuli using the direct magnitude estimation (DME) procedure with respect to a modulus, which was a nasal vowel synthesized with a moderate velopharyngeal opening area (i.e., VPO=$20mm^2$).

2.1.2 Preliminary results

The effect of articulatory adjustment on the acoustic characteristics of /i/

Figure 5.6 provides some examples of the spectra of O, N and NA samples. Figure 5.6(a)-(c), (d)-(f), and (g)-(i) correspond to three O/N/NA vowel sets and Figure 5.2 (a)-(c) provide the area functions for the three N/NA pairs in Figure 5.6. By comparing the spectral characteristics of the three vowels within a group, it was found that

1. In the nasal vowel (N) of the first example, the amplitude of F1 was reduced to some extent and the intensity of F2 was largely attenuated by the nasal zero below it, compared to the oral vowel in (a). In the nasal vowel with articulatory adjustment (NA), the amplitude of F1 stayed almost the same as the F1 of the oral vowel (O) in (a) and the intensity of F2 was preserved regardless of the adjacent nasal zero. In addition, the nasal zero in NA corresponded to a shallower spectral "valley" compared to the corresponding nasal zero in N.

2. In the second example, the F2 in N was attenuated by the adjacent nasal zero, whereas the F2 in NA was preserved due to an increased distance between F2 and the nasal zero.

3. In the third example, the frequency of F2 in the nasal vowel was shifted upward and the bandwidth of F2 became wider as a result of nasal coupling. The depth of the nasal zero below F2 was deeper compared to the corresponding nasal zero in (i). The effect of articulatory adjustment was recognized as to preserve the frequency and bandwidth of F2, as seen in (i).

The difference of area function between N and NA in Figure 5.2 marked out the critical articulatory adjustment(s) that lead to the corresponding acoustic changes as enumerated above. Specifically, in the first example, the articulatory adjustment resulted in an expansion of the vocal tract at the velopharynx and the upper pharynx and a constriction behind the lips, both changes marked with the arrows in Figure 5.2 (a). The second example showed prominent vocal tract expansions at the velopharynx, at/behind the lips after the articulatory adjustment, as marked with the arrows in Figure 5.2 (b). In the third example, the changes marked with the arrows in Figure 5.2 (c) indicated that the vocal tract was expanded at the velopharynx and behind the lips and the lip opening area stayed almost the same after the articulatory adjustment.

The effect of articulatory adjustment on the nasality rating of /i/

The raw nasality scores were equalized (see Appendix for equalization procedures) and then plotted for the three types of vowels in Figure 5.3. The median of the nasality scores for the nasal vowels with articulatory adjustment was shown to be lower than the median of the nasal vowels without adjustment and higher than the median of the oral vowels. The mean nasality score of the nasal vowels with articulatory adjustment is123.57 ($s.e.$ = 72.71), which is 21.30 lower than the mean nasality score of the nasal vowels without adjustment and 10.12 higher than the mean nasality of the oral vowels. The analysis of variance found a significant effect of the vowel type factor on nasality ratings ($F(2, 51) = 4.8323$, $p = 0.012$). Tukey's HSD post-hoc test showed that the mean nasality score of the nasal vowels with articulatory adjustment is significantly lowered ($p = 0.0441^*$) compared to the nasal vowels without articulatory adjustment, and, on the other hand, it is not significantly different from its oral counterpart ($p = 0.3314$).

2.1.3 Remarks on the pilot study

In general, the effect of articulatory adjustment on hypernasal speech includes: 1) compensating for the acoustic deviation caused by excessive velopharyngeal opening (e.g., attenuated/shifted F2, an extra nasal pole-zero pair at around 1000Hz) (Rong and Kuehn, 2010), and 2) reducing the perceived nasality (Rong and Kuehn, 2012). The articulatory model (Speech Processing and Synthesis Toolboxes) enables such articulatory adjustments by positioning and coordinating the articulators in such a way that minimizes the acoustic discrepancy between the model simulation and the target derived from an oral vowel. However, as the

sizes and positions of the hard structures (i.e., anatomical landmarks) are fixed in the model, individual anatomical variability cannot be accounted for. In terms of articulation, similar limitations related to the fixed articulatory space exist in the model. Therefore, the articulatory adjustment derived from the model is optimized in a sense that it maximally compensates for the acoustic effect of velopharyngeal opening for a "speaker" with the same parameters as the model, but it would not work in the same way for other speakers due to individual anatomical and articulatory variabilities. To accommodate such individual variabilities, a normalization is needed to adapt the standard articulatory model to individual speakers by taking into account inter-speaker anatomical and articulatory variabilities.

2.2 Development of a speaker-dependent articulatory model based on EMA and aerodynamic measures

The articulatory model (Speech Processing and Synthesis Toolboxes) used in the pilot study was modified to derive the articulatory adjustment that compensates for the acoustic and perceptual outcome of velopharyngeal opening in different individuals. As discussed in the previous sessions, to account for inter-speaker anatomical and articulatory variabilities, the model needs to be normalized/adapted. The model adaptation requires a large inventory of speech data with a variety of articulatory gestures covering the articulatory space of the speaker to serve as the training database, based on which the speaker-dependent anatomical dimensions and articulatory space are determined.

Point-wise articulatory data and audio signals were acquired simultaneously with the EMA AG500 system (Carstens, Germany) and a microphone from two American English speakers in Experiment I. The velopharyngeal opening area was estimated based on simultaneous recordings of electropalatography (Reading system, EPG 3) and speech aerodynamic signals using the hydrokinetic method in Experiment II. The nasal airflow was measured with a nasal mask, while the intraoral air pressure below the velum and the nasal pressure above the velum were measured by two flexible rubber tubes with one placed in the mouth and the other inserted into one of the nostrils. The velopharyngeal opening area was estimated based on the aerodynamic measures using the hydrokinetic method (Warren and Dubois, 1964). With the articulatory data (i.e., oropharyngeal articulatory positions and VPO), the standard vocal tract model was trained to adapt its PONM (i.e., length and transverse dimensions of the vocal tract) to fit the two individual speakers.

The EMA-based point-wise articulatory data were fitted into the model by 1) aligning the orientation and origin of the two coordinate systems (model-based coordinate system and EMA-based coordinate system) based on the positions of the anatomical landmarks (e.g., tragi, upper incisor), 2) normalizing the units of the

EMA coordinates by matching the length of hard palate between the speaker and the model and 3) adjusting the articulatory parameters of the model so that the estimated configuration of the vocal tract fit all the EMA articulatory positions. After such articulatory matching, the lowest three formant frequencies of the corresponding audio signal and the simulated speech of the model were compared. If mismatches emerge, the default PONM of the model in control of the vocal tract size and shape (i.e., length and transverse dimensions of the vocal tract) were adjusted in a trial-by-error way to minimize the acoustic discrepancy (i.e., weighted sum of the formant-frequency errors) between the model and the target. Such an adaptation enabled the model to account for inter-speaker anatomical variability. Meanwhile, the movement ranges of the articulators were derived from the speech inventory to compose a speaker-dependent articulatory space.

In this way, the normalization consisted of three steps, namely, articulatory fitting, anatomical adaptation, and articulatory adaptation, aiming at adapting the standard vocal tract model framework by Childers (2000) to fit the features of individual speakers. During the normalization, the first step made phoneme-dependent adjustments of the articulatory parameters in the model to match the point-wise EMA articulatory positions in the database following the inverse NM. The second step adjusted the PONM of the model to account for the individual anatomical variability. The third step customized the model with speaker-dependent articulatory features by adapting the space of articulatory movement of the model to individual speakers, given the precondition of a sufficiently-large database covering all possible articulatory gestures.

2.3 Experiment I: Speech articulatory, acoustic and aerodynamic data collection

In Experiment I, articulatory, acoustic and nasal aerodynamic data will be collected simultaneously using the AG500 articulography (EMA) system, a Countryman Isomax E6 directional microphone and a Scicon NM-2 nasal mask, respectively. The nasal airflow collected by the mask provides aerodynamic cues that mark out the start and end of nasalization, if occurs. For nasal tokens, such aerodynamic information will be used to assist segmenting the target stimuli (nasalized vowels) from the recorded signal. The articulography system records the movement of the freshpoints on the primary articulators. Given the recorded articulatory data as training speech samples, the PONM of the model were adjusted so that the acoustic characteristics of the simulated model output matched the acoustic features of the corresponding audio signal recorded simultaneously with the articulatory data. The adjustment of PONM was intended to adapt the anatomical dimension of the model to a speaker-dependent vocal tract. On the other hand, the movement ranges of the freshpoints defined the articulatory space of the speaker. Therefore, the purpose of Experiment I is to

provide the speech database for speaker-adaptation of the articulatory model.

2.3.1 Participants and speech materials

Two male speakers participated in this study, both of whom spoke American English as their first language. The speakers have normal speech and hearing, both are free from any oral-nasal anomalies and are between the ages of 18-36. Both of the speakers were absent of any sign of cold or upper respiratory infection at the time of participation and their speech was evaluated to be in a normal resonance condition by the experimenter.

Speech materials consisted of /CV/, /VC/, /CVC/ and /VCV/ syllables, where "V" included three corner vowels /a/, /i/ and /u/, "C" stood for stops, fricatives and nasal consonants with different places of articulation (i.e., /p/, /b/, /t/, /d/, /k/, /g/, /f/, /v/, /s/, /z/, /m/, /n/, /ŋ/). All consonants listed were included in all syllable positions with the exception of /ŋ/, which was only used in syllable final positions. These consonants vary in places and manners of articulation as well as voicing. Different consonantal contexts were included to 1) account for the coarticulatory effect of consonantal articulation on the vowels, and 2) explore the articulatory space of the speaker. The syllables were recorded within the following carrier phases: "Say ... again," "Say ... six times," "I said ... again" and "I said ... six times," depending on the syllable structure. The vowel portion was segmented from the syllable to serve as the target for model adaptation. The speech materials covered a variety of articulatory gestures including 1) gross tongue motion such as the front-back movement in the contrast of front versus back vowels (e.g., /i/ vs. /u/) and the upward-downward movement in the contrast of high versus low vowels (e.g., /u/ vs. /a/), and 2) well-defined tongue gestures such as the tongue tip raising-lowering movement in consonants /t/, /d/, /s/, /z/, /n/, tongue dorsum raising-lowering movement in consonants /k/, /g/, /ŋ/, and lip opening-closing in /p/, /b/, /m/, so the articulatory space of the speaker was well covered by exploring all speech stimuli. The coarticulatory effect introduced by the preceding and/or following consonant added articulatory versatility to vowel production. Nasal consonants, particularly, were coarticulated with the preceding/following vowels to simulate hypernasal speech in such way that velopharyngeal port stayed open during the partial or entire duration of the vowel.

2.3.2 Acoustic data acquisition

The acoustic signal was recorded through a Countryman Isomax E6 directional microphone set placed about 5cm from the corner of the mouth. The signal gain was modulated using an M-Audio Fast Track Pro preamplifier to an appropriate level so that the signal did not clip during the recording.

2.3.3 Articulatory data acquisition with EMA

Instrumentation

The articulatory data were collected with the EMA AG500 system (Carstens, Germany) at a sampling rate of 200Hz. The AG500 system is comprised of six transmitter coils housed in a clear Plexiglas cube with an internal dimension of $56.5cm * 52.7cm * 50.8cm$. AC current runs through six separate electromagnetic coils, each at a specific frequency, placed on the periphery of the cube. The six currents from these emitting electromagnets are induced to small electromagnetic receiver sensors when they are placed inside the cube. The induced current amplitudes of each of the six discrete frequencies are recorded by a central computer. Given the constant rate of electromagnetic amplitude decay ($1/r^2$, where r = the radius of the magnetic field around the emitter), the AG500 is capable of determining the distance of each sensor from each of the six emitting electromagnets. In this way, the AG500 can record the location of each sensor in a three-dimensional space at a sampling rate of 200 Hz. During recording, the dynamic movements of the articulators were tracked in three dimensions by the EMA sensors glued to the articulators.

System and subject preparation

The AG 500 system was turned on to warm up for at least one hour before recording. Twelve sensors, calibrated as an experimental set (to be discussed below), were coated with latex milk and left to dry out at least 12 hours before the experiment.

Before attaching sensors to the subject, an experimenter marked all the facial points where the sensors were attached using a surgical marker. This is to ensure the consistency of sensor placement across participants and, especially, in case of sensor falling/reattachment within a recording session. Eight sensors were attached to the articulators of the subject using surgical glue: 1) four sensors on the tongue, including one on the tongue tip (TT), anther on the tongue blade (TB: 0.5-1.0cm backward from TT), one on the posterior tongue dorsum (PTD: the deepest position where the sensor can be placed without causing physical discomfort of the participant), and one on the anterior tongue dorsum (ATD: halfway between TB and PTD), 2) two sensors on the teeth (UI: upper incisor and LI: lower incisor), and 3) two on the lips (UL/LL: upper/low border of vermilion in midline). Another sensor was glued to the laryngeal prominence (LP: also known as "adam's apple") to indicate the movement of the larynx. Three additional sensors were affixed to the bony structures (Nose: nose bridge, LT/RT: left/right tragus) to serve as reference landmarks for head movement correction. After the attachment of all EMA sensors, the subject was moved to the center of the EMA cube and all sensors were then plugged into the receiver by an experimenter. By attaching a grounding palate to the wrist, the subject was grounded to reduce the physiological noise from interfering

with the system. A microphone was placed around 5cm away from the subject's left lip corner to record the audio signal simultaneously with the speech articulatory movement.

2.3.4 Nasal aerodynamic data acquisition

Measurement of nasal pressure/airflow gives indirect cues of velar movement. The onset/offset of nasal flow can serve as objective measurements of the onset/offset of vowel nasalization across different tokens for a given speaker, something that cannot be easily derived from the acoustic signal alone (Shosted, 2009; Warren and Dubois, 1964). In order to measure nasal flow, participants wore a vented Scicon NM-2 nasal mask (Rothenberg, 1977). The mask was secured laterally using a Velcro strap running behind the ears and fastened at the back of the head; the mask was secured medially by a strap running from the top of the mask over the forehead and fastened at the back of the head. A tube (3 m long, 4 mm ID) was connected to the open outlet of the nasal mask on one end and a Biopac TSD160A (operational pressure +/- 2.5 cm H2O) pressure transducer on the other. The voltage output of the transducer was amplified by the connected Biopac DA100C differential bridge amplifier (gain = 5000) and passed to a Krohn-Hite 3360 analog filter, primarily as an anti-aliasing filter (10 kHz low-pass). The final filtered signal was routed through a National Instruments BNC-2110 shielded block connecter which provided serial connection to a National Instruments PCI-6013 data acquisition (DAQ) board installed in an HP xw4400 Workstation running Microsoft Windows XP (Version 2002, Service Pack 3). The signal was digitized at 1 kHz and recorded using custom-written scripts running in MATLAB 7.5.0 (R2007b) that accessed functions native to MATLABs Signal Processing Toolbox (V6.8).

2.3.5 System synchronization

Because the EMA and aerodynamic data are collected using two different systems, it is necessary to synchronize the output signals. EMA data were recorded as individual sweeps, and each sweep corresponded to a stimulus imbedded in a carrier phase. Aerodynamic data were recorded continuously throughout the entire session to avoid misalignment of the EMA/acoustic sweeps and aerodynamic sweeps due to operator errors. The Sybox-Opto4 unit included with the AG500 articulography provided time synchronization of the articulatory and acoustic data. This synchronization was performed automatically with the native Carstens recording software, using a positive voltage pulse at the beginning of each sweep and a negative voltage pulse at the end of each sweep. To synchronize the EMA/acoustic signals with the aerodynamic siganl, the signal carrying synchronizing pulses was split using a BNC Y-cable splitter and the duplicate signal was sent to the BNC-2110 (aerodynamic) data acquisition board. The sync signal was captured simultaneously with the

aerodynamic data (sampling rate 1 kHz). A MATLAB script automatically identified the time points of the pulses and parsed the aerodynamic data between them. These parsed aerodynamic signals were combined with the EMA and acoustic signals in MATLAB later.

2.3.6 System calibration

Both the EMA system and the aerodynamic instruments were calibrated independently. The AG500 uses a proprietary calibration software created by the manufacturer of the articulography. Twelve sensors are calibrated together as an experimental set, and all sensors in a set are recalibrated when one or more sensors need to be replaced due to wear. A calibration file is only valid for a specific sensor set. During the calibration, twelve sensors are mounted to a machined cylinder-and-plate device known as a circal. The placement of the sensors on the circal suspends them in the center of the cube, and the AG500 rotates the circal 360 degrees. During this rotation, the 3D position, tilt, and yaw of each sensor with relation to the six electromagnetic emitters are recorded. The AG500 system later uses this information to calculate the positions of the sensors by converting the voltage amplitude of each of the six frequency-distinct electromagnetic fields into a position relative to the emitters. The calibration session file can be used multiple times with the same set of sensors, until the sensor set requires recalibration.

The aerodynamic system was calibrated before each recording session for each speaker. The Scicon NM-2 mask (connected to a Biopac 160A transducer and to the BNC 2110 data acquisition interface as described above) was held against a custom-designed plaster negative of the mask, creating an airtight seal. A tube ran from a hole drilled in the plaster negative to a Boxer 7004 quad-headed gas pump. The pump generated an outflow of 1033 ml/s and an inflow of -1033 ml/s with a pause (0 ml/s) between pulses. Positive and negative pulses were recorded separately. The electrical response of the transducer was measured individually for the positive pulses, negative pulses, and flat values (zeros). These were averaged to produce single values for positive pulses (1033 ml/s), negative pulses (-1033 ml/s) and zeros (0 ml/s). The electrical response of the transducer was plotted against the known value of the flow pulses and a linear function was fitted to the data points. The coefficients of this function were used to define the calibration function, which was used to transform the raw electrical output of the transducer.

2.3.7 Recording procedures

Prior to speech recording, the subject was asked to talk at a conversational level with the EMA sensors and the nasal mask attached so that he can get accustomed to the systems. Meanwhile, an experimenter adjusted the gain of the amplifier to allow for an appropriate speech intensity level. Recording started when

the experimenters perceived the speech of the subject to sound sufficiently natural. During the recording, the subject was asked to read the speech stimuli displayed on a series of PowerPoint slides on a laptop monitor. An experimenter sat three feet in front of the cube and advanced the slides after the subject articulated each stimulus. At the end of recording, one experimenter took off a sensor from the subject and glued it to one of her fingers with the gloves on to trace the mid-sagittal contour of the hard palate for three times. Three experimenters were present during each recording session: one of them presented speech materials to the speaker; one controlled the EMA recording and monitored the real-time display of sensor movement to detect potential sensor errors (e.g. sensor failure, falling, etc.); and the other controlled and monitored the aerodynamic recording.

2.4 Experiment II: Estimation of velopharyngeal opening aperture using the hydrokinetic method

In Experiment II, linguopalatal contacts, acoustic and pressure/flow aerodynamic signals were recorded using the Reading EPG 3 system, a head-mounted AKG C520 cardioid microphone and a Gold Seal nasal CPAP mask plus two flexible plastic tubes, respectively. The same two subjects as in Experiment I participated as speakers. The speech stimuli included all the nasal tokens in Experiment I and were recorded three times by each speaker.

The pressure-flow signals were used to estimate the velopharyngeal opening area based on the hydrokinetic method. As EMA does not provide articulatory information on the velar movement, the estimation of VPO area from the physiological measures completes the articulatory information necessary to model adaptation.

2.4.1 Acoustic data acquisition

Speakers wore a head-mounted AKG C520 cardioid microphone (Harman International, Stamford, CT). The audio signal was passed to a Grace m101 pre-amplifier (Grace Design, Boulder, CO) and then to one of two line inputs of the WinEPG EPG3 Serial Interface SPI V2.0 (SPI).

2.4.2 EPG-aerodynamic data acquisition

Instrumentation

Data were recorded using the Articulate Instruments (AI) software interface (v. 1.17; Articulate Instruments, Musselburgh, UK). Each speaker was recorded while wearing a thin acrylic electropalate manufactured by Incidental (Newbury, Berkshire, UK). Each electropalate is designed using a unique superior maxillary model

cast by a local, board-certified orthodontist. Each electropalate has 62 silver electrodes arrayed across eight rows proceeding from front to back; the anterior-most row has six electrodes and the rest have eight. The electropalate is secured to the roof of the speaker's mouth with wire clasps mounted on the palate; these fit snugly around the 1st or 2nd molar, depending on the speaker's dentition. The palate is connected to the WinEPG multiplexer, to the WinEPG Palate Scanner EPG3.V2, then to the WinEPG EPG3 Serial Interface SPI. Whenever the tongue touches an electrode, an electrical signal is sent through this chain. Given that the position of each electrode is known, the electrode signals are ultimately interpreted as indicating whether a specific region of the palate is in contact with the tongue. The sampling rate of the scanner is 100 Hz.

In order to completely assure the safety of the subjects using the EPG, every electrical device that the pseudo-palate wearer can reach or is connected to, such as the PC and computer monitor, are isolated from the main power supply through a medical isolation transformer that has been designed to the IEC 60601-1 standard.

The intraoral air pressure below the velum, the nasal air pressure above the velum and the nasal airflow were acquired at 100Hz simultaneously with the EPG data. To record intraoral air pressure below the velum, an experimental setting was designed whereby a flexible rubber tube was affixed to the posterior end of the EPG palate on one end, and connected to a Biopac TSD160B pressure transducer on the other end. The voltage output of the transducer was amplified by the Biopac DA100C differential bridge amplifier, passed to the UIM100C interface modulus, and recorded by the AcqKnowledge software. This setting allowed for simultaneous recordings of linguopalatal contact and intraoral pressure, and the subject was more comfortable compared to having a tube or catheter inserted to the velopharynx through the nose.

Nasal airflow was recorded with a Gold Seal nasal CPAP mask (Philips Respironics, Eindhoven, The Netherlands). The mask was held and pressed against the nose to form an air-tight seal by the subject. The vent of the nasal mask was plugged with a heated Fleisch Pneumotach oriented so that air from the mask was channeled through the pneumotach (Biopac TSD137; Biopac Systems, Goleta, CA). The pneumotach pressure ports were connected via rubber tubing to a Biopac TSD160B pressure transducer ($\pm$ 12.5 cm H_2O). The voltage output of the transducer was amplified by the Biopac DA100C differential bridge amplifier (gain = 5000). To measure nasal pressure together with airflow, a hole was drilled on the side of the mask and a rubber tube was fitted into the hole and inserted into the less resistant nostril of the subject. The end of the tube went as far inside the nasal cavity as not causing physical discomfort of the subject. The other end of the tube was connected to a Biopac TSD160B pressure transducer. The voltage output of the transducer (TSD160B) was amplified by the Biopac DA100C differential bridge amplifier, passed to the UIM100C modulus and recorded by AcqKnowledge.

2.4.3 System synchronization

Because EPG/acoustic and pressure/flow data were recorded by different acquisition systems, a system-synchronization was needed. This was enabled by an external trigger that sent out a negative pulse when triggered and a positive pulse when released. The pulse signal was split into two channels, amplified, and passed into two systems: one to the line input of the WinEPG EPG3 Serial Interface SPI and the other passing through UIM100C to AcqKnowledge. An experimenter controlled the trigger to start and end the recording of each sweep. EPG/acoustic and aerodynamic signals were aligned based on the synchronization pluses.

The audio and pulse signals were passed from the SPI to an AudioFire2 (Echo Digital Audio, Carpinteria, CA) IEEE 1394 serial bus interface for isochronous real-time data transfer to an HP xw4400 Workstation running Microsoft Windows XP (Version 2002, Service Pack 3). The EPG signal was passed from the SPI to the same computer via USB. Synchronization of the audio, synchronization pulses, and EPG signal is a function of the WinEPG hardware and Articulate Instruments 1.17 software installed on the machine. Successful synchronization depends to some extent on the audio transmission hardware (e.g. sound card). According to Alan Wrench (p. c. June 23, 2009), "The poorest sound card might drift by one frame (10 ms) after six seconds of recording". Recordings for this experiment were considerably shorter. In addition, a high-speed AudioFire2 interface was used to transmit the audio and pulse signals directly to the computer, bypassing its internal sound card completely. The pulse and EPG signals were regarded to be synchronous. Audio and synchronization pulses were both digitally sampled at a rate of 48000Hz.

2.4.4 System calibration

Nasal flow was calibrated using a Boxer 7004 quad-headed gas pump (Uno International, London) that generated positive and negative flow of 1033.33 ml/s. The nasal mask, fitted with the pneumotach, was attached to a custom-designed plaster cast connected to the pump. When the pump works, the Biopac TSD160B pressure transducer connected to the pneumotach pressure port registered the pressure generated by the pump in an electrical form.

Pressure signals were calibrated using a water manometer (Respironics Inc., Murrysville, PA), which measured the pressure difference between the atmosphere and the calibration pressure by balancing the weight of a fluid column. The water manometer was connected to the Biopac TSD160B pressure transducer via a plastic tube to transfer the calibration pressure signal to TSD160B.

The electrical response of the transducer was plotted against the known value of the calibration pressures and a linear function was fitted to the data points. The coefficients of this function defined the calibration

function, which was used to transform the raw electrical output of the transducer into the aerodynamic form.

2.4.5 Recording procedures

Speaker wore the EPG palate with the electrode wires and an oral tube passing through the buccal cavity to outside of the mouth. The ends of the wires and tube were connected to the corresponding acquisition systems. The nasal mask was held and pressed against the nose by the subject to ensure an air-tight seal. The nasal tube was inserted through the less resistant nostril into the nasal cavity. The open ends of the nasal mask and nasal tube were connected to the corresponding acquisition system. When the subject was ready, the experimenter started the EPG acquisition system (Articulate) and the aerodynamic acquisition system (AcqKnowledge) to initiate recording. The trigger was controlled by the experimenter to begin and end recording of each EPG/acoustic sweep, whereas the aerodynamic signal was recorded as a long sweep containing all stimuli.

2.5 Data processing

2.5.1 EMA normalization for head movement correction

A normalization procedure was developed to decouple the speech-unrelated head movement from the speech-related articulatory movement in the kinematic signal measured by the EMA sensors. The normalization based on a coordinate transformation was intended to calculate the sensor positions within a head coordinate system so that the head-movement component was excluded from the articulatory movement tracked by the EMA sensors.

There were four steps in the normalization: 1) a rotational transformation was applied to adjust the orientation of the head so that the vector pointing from the midpoint of the left and right tragi (MT) to the nose bridge (Nose) (i.e., MT-Nose vector) was rotated to the midsagittal plane with an angle of 0.3412 from the x-axis (the angle was determined by the parameters of the standard model framework by Childers (2000)); 2) a translational transformation was applied to shift the location of MT to the origin (i.e., [0, 0, 0]); 3) with the MT-Nose vector as the rotational axis and the position of MT as the origin, the head was rotated across the axis with the origin fixed so that the vector connecting LT and RT was rotated to the transverse plane; and 4) the matrices that represented the transformations in steps 1)–3) were applied to the positions of all 12 sensors so that the transformed sensor positions all had the same reference frame, which was the speaker's head. Such a normalization was applied to each recording sweep of EMA so that the speech-unrelated head movement was excluded from the articulatory movement represented by the sensor

positions in the head-based coordinate system.

2.5.2 Annotation

EMA, acoustic and nasal-aerodynamic data annotation

Based on the acoustic signal, each token was marked manually with three landmarks, corresponding to the onset, midpoint and offset of the vowel portion. Vowel onset was determined as the onset of regular vibration in the sound pressure waveform. Accordingly, vowel offset was selected as the offset of regular sound wave vibration, characterized by the cession of voicing or dramatic amplitude change in the waveform. Given the onset and offset, the midpoint of the vowel was marked in accordance. Figure 5.4 gives an example of the nasal token (/baŋ/) to illustrate the annotation scheme, where the two vertical solid red lines mark the onset and offset of the vowel.

For nasal tokens, two additional landmarks related to nasalization were marked based on the nasal airflow signal. The first derivative of the low-pass filtered nasal flow signal was first calculated. A threshold at 20% above the average filtered nasal flow velocity was set for each sweep. In anticipatory nasalization, the onset of nasalization was determined as the first positive velocity peak above the threshold that occurred after the voice onset of the vowel. The midpoint between the onset of anticipatory nasalization and offset of vowel was marked as the midpoint of nasalization. In carry-over nasalization, the offset of nasalization was determined as the last positive velocity peak above the threshold before the vowel offset. The midpoint between the onset of vowel and the offset of carry-over nasalization was marked as the midpoint of nasalization. If both anticipatory and carry-over nasalization occurred in a vowel, the portion with the nasal airflow velocity above the threshold was marked to be the nasalized part. If the nasal airflow velocity never went above the threshold along the vowel duration, the vowel was regarded as denasalized. On the other hand, if the nasal airflow velocity was always above the threshold during the vowel portion, the vowel was considered as fully nasalized. In Figure 5.4, the vertical dashed line marked out the onset of anticipatory nasalization, which corresponded to the first velocity peak of nasal airflow after the onset of the vowel.

From an articulatory view, the following landmarks related to the vertical movement of the primary articulator were annotated in addition to the acoustic landmarks: 1) the maximum vertical position of TT for /t/, /d/, /n/, /s/, /z/; 2) the maximum position of PTD for /k/, /g/, /ŋ/; and 3) the maximum and minimum vertical distance between UL and LL for /p/, /b/, /m/. Furthermore, within each vowel segment determined by its acoustic boundaries, the maximum vertical position of PTD was marked for high vowels /i/ and /u/ and the minimum vertical position of PTD was marked for the low vowel /a/. In addition, the minimum vertical position of the primary articulator was marked for each vowel. Figure 5.5 gives a graphical

illustration of the articulatory annotation scheme using the word "dike" as an example.

The vowel segments (oral and nasalized) determined by both acoustic and articulatory landmarks were used to serve as the training speech samples for adaptation of the PONM of the model. The aerodynamic landmarks accounted for the timing of nasalization (i.e., beginning, midpoint and end of nasalization). The articulatory landmarks related to the consonant production designated critical articulatory gestures within each stimulus that, taken together, covered the articulatory space of the speaker. Therefore, the articulatory information at these landmarks was used to compose the database to compute the speaker-dependent ranges of articulatory movement.

Pressure-flow data annotation

The peaks of nasal airflow were marked and the intraoral and nasal pressures at these peaks were measured. Based on the pressure-flow measures, the velopharyngeal opening area was estimated at these annotated nasal-airflow peak. In addition, the start, midpoint and/or end of nasalization were marked based on the nasal airflow and acoustic signals according to the annotation scheme above. The velopharyngeal opening area was estimated at these time frames as well.

2.5.3 Formant-frequency measurement

At the annotated time frames within each vowel segment, the lowest three formant frequencies of the vowel were computed using the Linear Predictive Coding (LPC). A Hamming window was applied to each vowel sample with its center at the time point of interest and a length of 512. Using the FFT function in MATLAB R2011_a and the FREQZ function in MATLABs Signal Processing Toolbox,16th-order Linear Predictive (LP) filters were designed for oral vowel samples and 30th-order filters were designed for nasal vowel samples. Peaks in the LP filter were detected automatically in MATLAB. For oral vowels, an order of 16 was selected to estimate approximately eight formants in the frequency range 0–8 kHz. For nasal vowels, a higher order was chosen because nasal vowels are expected to have more peaks in their spectra (especially in the low-frequency region) due to coupling of the nasal tract, asymmetry of nasal passages, and branching of paranasal sinuses. Because LPC is known to be precise but not robust in detecting spectral peaks, some spurious peaks/formants, usually characterized by wide bandwidth and small amplitude, could be detected. An arbitrary bandwidth threshold of 200 Hz was set to exclude such peaks.

Figures of each 2048-point FFT spectrum with a LPC-based envelope were generated. The means (Mean) and standard deviations (SD) of the lowest three formant frequencies (F1, F2, F3) detected by LPC were calculated. Whenever any of the three formant frequencies fell out of the range of $(Mean - 2 * SD)$ –

($Mean + 2 * SD$), the corresponding formant frequency was logged interactively in MATLAB by manually measuring the frequency of harmonic closest to the formant peak in the spectrum of the vowel sample. For nasal vowels, such a false detection of formant happened more frequently due to the presence of both additional formants and anti-formants in the spectra.

After checking of all LPC-detected formants, the lowest three formant frequencies of all vowel samples after manual correction served as the acoustic targets for model adaptation.

2.5.4 VPO estimation

The VPO area was estimated from the aerodynamic measures recorded in Experiment II. At each selected time frame, the nasal pressure above the velum (P_n), intraoral pressure below the velum (P_o) and nasal airflow rate ($\dot{V}_n$) were used in the hydrokinetic equation as follows to calculate the velopharyngeal opening area:

$$VPO = \frac{\dot{V}_n}{k\sqrt{2(\frac{P_o - P_n}{D})}} \tag{2.1}$$

where VPO is the estimated opening area in cm^2, D is the density of air and k is a correction factor (estimated as 0.65 by Warren and Dubois (1964)).

2.6 Model fitting and adaptation

2.6.1 Articulatory fitting

The articulatory information at eight time frames of nasal tokens (i.e., onset, midpoint and offset of the vowel; onset, midpoint and offset of nasalization; minimum vertical position of the primary articulator; maximum/minimum vertical position of PTD for the high/low vowel), and at five time frames of oral tokens (i.e., onset, midpoint, and offset of the vowel; minimum vertical position of the primary articulator; maximum/minimum vertical position of PTD for the high/low vowel) served as the input for adaptation of the PONM of the model.

In the framework of the model by Childers (2000), the entire vocal tract is divided into six sections from the vocal folds to the lips, including two pharyngeal sections (AR1, AR9) extending from the vocal folds to the velopharynx, one posterior oral section (AR2) covering the region from the velopharynx to the posterior dorsum of the tongue, one main oral section (AR23) covering the oral cavity from the posterior tongue dorsum to the tongue tip, one front oral section (AR4) covering the space between the tongue tip and the upper

and lower incisors, and one lip section (AR5) extending from the incisors to the lip opening. A graphical representation of Childers's vocal tract model is shown in Figure 5.7. The midsagittal configuration of each section is determined by the placement of the relevant articulator(s) and the anatomy of the hard structures (e.g., posterior pharyngeal wall, hard palate, etc.). Therefore, the first step was to fit the articulatory positions at the selected time frames into the model framework to estimate the midsagittal shape of the vocal tract for each vowel sample.

First of all, the midsagittal trace of the hard palate contour was fitted into a 2D standard vocal tract model framework (Childers, 2000) by aligning the position of the upper incisor (UI) as an anatomical landmark. The midsagittal positions of the other eleven sensors on TT, TB, ATD, PTD, LI, UL, LL, Nose, LT, RT, and Larynx were determined accordingly within the model framework. Among all the EMA sensors, the positions of TT, TB, ATD, PTD, LI, UL and LL, taken together with the estimated VPO area based on the hydrokinetic method, were used to determine the shape of the vocal tract.

According to the graphical representation of Childers's vocal tract model in Figure 5.8, the configuration of the vocal tract is determined by the positions of a set of articulators. As EMA did not provide measurements on the pharyngeal movement, the default hyoid positions for the three vowels (i.e., /a/, /i/, /u/) given in Childers (2000), taken together with the position of PTD, were used to estimate the inner surface shapes of AR1, AR9 and AR2. Specifically, the center (*tongc*) and radius (*tongr*) of the posterior part of the tongue body were determined by the positions of PTD and ATD so that the midsagittal contour of the tongue surface between PTD and ATD composed a $\pi/2$ arc with its center at *tongc* and a radius of *tongr*. Given the position of *tongc*, the landmark (DL) that separated the upper and lower pharynx was determined such that the vector pointing from *tongc* to DL had a length of *tongr* and was perpendicular to the vector pointing from the neutral position of the hyoid bone (H) to DL. The locations of H and DL served as the lower and upper boundaries of the lower pharyngeal section (AR1), where the midpoint (PP) of AR1 was determined by the deviation of the hyoid bone for each vowel from its neutral position. Given the positions of H, PP and DL, the midsagittal shape of inner surface of the lower pharyngeal section (AR1) was simulated as an arc passing H, PP and DL with its center (*pme*) at the crossover of two vectors v1 and v2, where v1 was a vector perpendicular to the DL-PP line with its origin at the midpoint of DL and PP; v2 was a vector perpendicular to the PP-H line with its origin at the midpoint of PP and H.

The midsagittal inner surface of a combination of the upper pharyngeal section (AR9) and posterior oral section (AR2) was simulated as an arc passing DL and PTD with a center at *tongc* and a radius of *tongr*. For the main oral cavity (AR23), its inner surface configuration was determined by the four EMA freshpoints on the tongue (PTD, ATD, TB, TT). The part between PTD and ATD was simulated as an arc with its center at *c1b* and a radius of *r1b*, where *c1b* was located on the perpendicular bisector of the PTD-ATD line with a

distance of 1.5 times of the distance between PTD and ATD (i.e., $1.5||PTD-ATD||$) to the PTD-ATD line. Similarly, the portion between ATD and TB was simulated as an arc with its center at *c2b* and a radius of *r2b*, where *c2b* was located on the perpendicular bisector of the ATD-TB line with a distance of 1.5 times of the distance between ATD and TB (i.e., $1.5||ATD-TB||$) to the ATD-TB line. The part between TB and TT was simulated as another arc with its center at *c1t* and a radius of *r1t*, where *c1t* was located on the perpendicular bisector of the TB-TT line with a distance of 1.5 times of the distance between TB and TT (i.e., $1.5||TB-TT||$) to the TB-TT line. The midsagittal shape of inner surface of the anterior oral section (AR4) was simulated as a straight line connecting TT and PF, where PF was a static structure with its position linearly controlled by the movement of LI. The last section AR5 was simulated as a straight tube with a constant midsagittal distance equal to the vertical distance between UL and LL.

The EMA freshpoints, taken together, determined the inner contour of the vocal tract in the midsagittal plane. On the other hand, the outer surface of the vocal tract was mostly static except the AR2 and AR5 sections. For the posterior oral section AR2, the outer contour was simulated as an arc passing the uvula (V) and the posterior end of the hard palate (M), where the center (*cmn*) of the arc was on a vertical line passing M so that the distance between *cmn* and M equalled the distance between *cmn* and UI and such a distance was also the radius (*rc*) of the M-V arc. For the lip section AR5, the outer contour was determined by the vertical position of UL. Apart from AR2 and AR5, all other sections had static outer contours. For AR1 and AR9, the outer contours were simulated to follow the framework of Childers's model. For the main oral cavity AR23 and the anterior oral cavity AR4, the outer contour of these two sections followed the midsagittal trace of the hard palate measured by EMA.

After articulatory fitting, both the outer and inner contours of the vocal tract were determined in the midsagittal plane, which composed a midsagittal profile of the vocal tract configuration to provide the source for model adaptation in the next step.

2.6.2 Adaptation of PONM

With the midsagittal shape of the entire vocal tract determined for each vowel sample based on the procedures above, default values were assigned to the PONM of the model to initiate the adaptation of PONM. The PONM, which determine the length and transverse dimension of the vocal tract, were composed of eleven independent parameters, including two parameters (coef_phar, coef_oral) in control of the lengths of the pharyngeal and oral cavities respectively, and seven parameters (slp1, int1, slp2, slp23, slp4a, slp4b, slp4c, slp5, int5) in control of the transverse dimension of the vocal tract.

As the entire vocal tract was divided into 60 consecutive tubes, the shape of the vocal tract can be

represented by the area and length functions as two 60-by-1 vectors. To compute the area and length functions requires the midsagittal articulatory configuration as determined above and the eleven PONM parameters. The parameter coef_phar is the ratio of the pharyngeal cavity length (i.e., sum of the first half elements of the length function) to the default length of the pharyngeal cavity. Similarly, coef_oral is the ratio of the oral cavity length (i.e., sum of the second half elements of the length function) to the default length of the oral cavity. The other nine PONM parameters convert the midsagittal distance to the area function using the following equations:

For AR1 and AR9,

$$A = \frac{\pi}{2} * (slp1 * (gc/gw) + int1) * d * cf \tag{2.2}$$

where A is the cross-sectional area of a tube within the sections of AR1 and AR9, gc and gw are two constants related to the geometry of the pharynx, d is the midsagittal distance, and cf is the cosine of the angle between the cross-section of the tube and the midline of the vocal tract at the corresponding location.

For AR2,

$$A = slp2 * d^{1.5} * cf \tag{2.3}$$

where A is the cross-sectional area of a tube within the AR2 section, d is the midsagittal distance, and cf is the cosine of the angle between the cross-section of the tube and the midline of the vocal tract at the corresponding location.

For AR23,

$$A = slp23 * d^{1.5} * cf \tag{2.4}$$

where A is the cross-sectional area of a tube within the AR23 section, d is the midsagittal distance, and cf is the cosine of the angle between the cross-section of the tube and the midline of the vocal tract at the corresponding location.

For AR4,

$$A = cf * \begin{cases} slp4a * d & d < 0.5 \\ slp4b * (d - 0.5) + 0.5 * slp4a & 0.5 \leq d \leq 2 \\ slp4c * (d - 2) + (1.5 * slp4b + 0.5 * slp4a) & d > 2 \end{cases} \tag{2.5}$$

where A is the cross-sectional area of a tube within the AR4 section, d is the midsagittal distance, and cf is the cosine of the angle between the cross-section of the tube and the midline of the vocal tract at the corresponding location.

For AR5,

$$A = \frac{\pi}{2} * [d * (slp5 * (sl - pl) + int5)] * cf \tag{2.6}$$

where A is the cross-sectional area of a tube within the AR5 section, sl and pl are two parameters related to lip protrusion and opening, d is the midsagittal distance, and cf is the cosine of the angle between the cross-section of the tube and the midline of the vocal tract at the corresponding location.

The adaptation of PONM used the simulated annealing algorithm with the default values of PONM as the initial input and the acoustic discrepancy between the model simulation and the target as the error term of the evaluation function.

$$error = 0.4 * \frac{||\tilde{F}1 - F1||}{F1} + 0.4 * \frac{||\tilde{F}2 - F2||}{F2} + 0.2 * \frac{||\tilde{F}3 - F3||}{F3} \tag{2.7}$$

where $F1$, $F2$, $F3$ are the lowest three formant frequencies measured from the target vowel and $\tilde{F}1$, $\tilde{F}2$, $\tilde{F}3$ are the lowest three formant frequencies computed from model simulation.

The PONM were adjusted in a trial-by-error manner so that during each trial, the acoustic error in Equ. 2.7 was reduced compared to the last trial and the iteration of adjustment stopped when the error was minimized or when the number of iterations reached its upper limit (i.e., 10000).

The optimization of PONM was applied to each vowel sample to adjust the length and transverse dimension of the model to fit a speaker-dependent vocal tract. With a set of optimized PONM computed for all vowel samples from a speaker, the average of each PONM parameter was calculated across all samples to compose the adapted PONM set for the speaker.

2.6.3 Adaptation of the articulatory space

In addition to the adaptation of PONM, the articulatory space of the model was also adapted to individual speakers by adjusting the ranges of the controlling articulatory parameters. Specifically, after the midsagittal configuration of the vocal tract was determined for each speech sample using the articulatory fitting process as described above, the controlling parameters of the model, including "hyoid bone height" (H), "velic opening" (VPO), "tongue body position" (*Tbody-x* and *Tbody-y*), "tongue tip position" (*Ttip-x* and *Ttip-y*), "jaw position" (*Jaw-x* and *Jaw-y*), "lip protrusion" (Lpro) and "lip opening" (*Lopn*), were adjusted to fit the estimated midsagittal contour of the vocal tract. Such an adjustment of the controlling articulatory parameters (ART=$\{H$, *Tbody-x*, *Tbody-y*, *Ttip-x*, *Ttip-y*, *Jaw-x*, *Jaw-y*, Lpro, *Lopn*$\}$) was applied to each speech sample within the database (including both vowel and consonant segments) to result in a set of ART that contained all possible articulatory positions of the speaker. The maximum and minimum values of each ART parameter were calculated to serve as the upper and lower boundaries of the movement range of each articulator. In this way, the articulatory space represented by the movement ranges of all articulators was adapted to individual speakers.

2.7 Using articulatory adjustment to compensate for the acoustic outcome of VPO

2.7.1 Articulatory adjustment

With the speaker-dependent PONM and articulatory space, the articulatory model was regarded to be speaker-adaptive and was used to develop speaker- and speech-based articulatory adjustment to compensate for the acoustic outcome caused by excessive VPO.

The articulatory adjustment was made by coordinating the articulatory parameters of the model (i.e., ART) customized with speaker-dependent PONM and articulatory space to achieve the acoustic goal defined by a target vowel sample using the simulated annealing algorithm. Specifically, the target vowels were selected as all of the oral and nasal vowel samples in the database, whose formant frequencies F1, F2 and F3 were measured to serve as the acoustic target for articulatory adjustment. The PONM of the model were assigned with speaker-dependent values computed during the PONM adaptation as the means of all the optimized PONM parameters across all speech samples of the speaker. Given the PONM values and the acoustic target, the simulated annealing algorithm adjusted the articulatory parameters of the model (ART={*H*, *Tbody-x*, *Tbody-y*, *Ttip-x*, *Ttip-y*, *Jaw-x*, *Jaw-y*, Lpro, *Lopn*}) in a trial-by-error manner to minimize the error term of the evaluation function defined in Equ. 2.7. Based on an oral vowel target, two types of articulatory mappings corresponding to two different types of vowels (i.e., oral vowels and nasal vowels with articulatory adjustment) were generated to simulate the corresponding articulatory configurations. Based on a nasal vowel target, an articulatory configuration corresponding to a nasal vowel was simulated. Specifically, the oral vowel (O) configuration was simulated by first setting $VPO = 0$ and then adjusting the ART of the model within the speaker-specific ranges computed during the articulatory space adaptation to achieve the corresponding oral acoustic target. The nasal vowel (N) configuration was simulated by first setting $VPO = 200mm^2$ (the value was determined by taking into account both the estimated VPO area in Experiment II and a perceptual evaluation by the experimenter) and then adjusting the ART to achieve the corresponding nasal acoustic target. The articulatory configuration of a nasal vowel with articulatory adjustment (NA) was simulated by setting $VPO = 200mm^2$ and then adjusting the ART of the model to achieve the same acoustic target as the one used for simulation of O. In this way, the articulatory configurations of O, N and NA simulated oral speech, hypernasal speech and adjusted hypernasal speech, respectively.

2.7.2 Articulatory synthesis

Similar to the pilot study, three types of vowels (i.e., O, N, NA) were synthesized based on the articulatory configurations generated above. Given an excitation source with a fundamental frequency frequency of $f_0 = 120Hz$, a duration of 1 second, a gain of 60 dB, and the default settings of the LF model of the glottal waveform in Childers (2000), a filtering function determined by an articulatory configuration generated above was used to synthesize the corresponding vowel sample.

A total of 198 vowel samples (66 O + 66 N + 66 NA) including /a/, /i/ and /u/ were synthesized with the articulatory model for Speaker 1 and 186 vowel sample (62 O + 62 N + 62 NA) were synthesized with the articulatory model for Speaker 2.

2.8 Decomposition of orthogonal articulatory modes from area function

To compare the oropharyngeal articulations of the oral, nasal and adjusted nasal vowels, the area functions of O, N, NA were computed. The area functions of N and O were then used to decompose a series of orthogonal modes responsible for the primary oropharyngeal articulatory distinction between N and O. Similarly, the area functions of NA and O were used to decompose a series of orthogonal modes responsible for the primary oropharyngeal articulatory distinction between NA and O. The decomposition followed a procedure similar to Story and Titze (1998), and can be represented by the following equations:

$$\alpha(v, s) = A(v, s) - A_0(v) \tag{2.8}$$

where $v = 1, 2, 3$ corresponding to the three vowels /a/, /i/ and /u/ respectively, $A(v, s)$ is the area function of the s^{th} N or NA vowel sample and $A_0(v) = \frac{1}{M}\sum_{s=1}^{M} A(v, s)$ is the average of the area functions of all the M O vowel samples.

$$R_{ij} = \frac{1}{M-1}\sum_{s=1}^{M} \alpha(v_i, s)\alpha(v_j, s) \tag{2.9}$$

where $i, j = 1, 2, ..., N$ with $N = 60$ corresponding to the length of the area function vector.

$$R\phi = \phi I\lambda \tag{2.10}$$

where R is the covariance matrix in Eq. 2.9, I is the identity matrix, ϕ and λ are the empirical orthogonal modes and their corresponding eigenvalues, respectively.

$$pp(i,s) = \sum_{j=1}^{N} \alpha(v_j, s)\phi_i(v_j) \tag{2.11}$$

where $i = 1, 2, ..., N$ corresponding to the i^{th} mode and $pp(i,s)$ is the amplitude coefficient computed as the projection of the s^{th} α vector on the i^{th} mode.

In this way, the oropharyngeal articulatory difference between nasal and oral vowels can be represented by the empirical orthogonal modes (ϕ) of N. The oropharyngeal articulatory difference between adjusted nasal vowels and oral vowels can be represented by the empirical orthogonal modes (ϕ) of NA. The amplitude coefficients (pp) indicate the amount of variance of the nasal-oral oropharyngeal articulatory difference (i.e., N-O and NA-O) that is accounted for by each orthogonal mode. By comparing the orthogonal articulatory modes of NA and N, the articulatory adjustment made by the model can be examined straightforwardly to indicate clinical implications.

2.9 Experiment III: Rating of nasality

2.9.1 Participants and perceptual stimuli

Eleven subjects (3 males and 8 females) participated in the experiment, all of whom spoke American English as their first language. The participants all have normal speech and hearing and are between the ages of 18-30. All of the participants were absent of any sign of cold or upper respiratory infection at the time of participation.

The perceptual materials included 384 synthetic vowels (138 /a/ + 156 /i/ + 90 /u/) consisting of three types (O, N and NA).

2.9.2 Listening task

The stimuli were presented to the listeners in random orders. The listeners were asked to rate the nasality scores of all stimuli using the method of Direct Magnitude Estimation (DME) with modulus. Each stimulus was presented to the listener following a modulus, which is a nasal vowel synthesized with the articulatory synthesis function based on the same acoustic target as the stimulus with a moderate VPO of $20mm^2$. Each stimulus was paired with its modulus so that the listener rated the nasality of the stimulus with respect to the modulus, which was assigned an arbitrary nasality score of 100.

Before the listening task, a training session consisting of 10 synthetic vowel samples (including /a/, /i/, /u/ and O, N, NA), not drawn from the perceptual stimuli for the listening task and synthesized with the same articulatory synthesis function, was provided to the listeners with their corresponding moduli so that the listeners can familiarize themselves with the DME procedures and the perceptual quality of synthetic vowels through practice. The listeners were asked to practice with the training vowel sample as many times as they want until they felt ready to proceed to the listening task.

During the listening task, the listeners rated the nasality score of each stimulus with respect to the preceding modulus using the DME method, which sets no upper/lower boundaries of the score and assumes a prothetic scale. In other words, the listeners were asked to use whatever number that seemed appropriate to them to quantify the degree of nasality. For example. if the stimulus sounded eight times as nasal as the modulus, then a nasality score of 800 should be assigned to the stimulus; if the stimulus sounded half as nasal as the modulus, then the nasality score should be 50. Each stimulus was played once to the listeners. If the listeners have a difficulty in making an immediate judgement after the presentation of the stimulus, they have a chance to replay the stimulus once.

Chapter 3

Results

3.1 Articulatory fitting

Figure 5.9 – 5.17 give 18 examples of the midsagittal configurations of the vocal tract estimated from the EMA measurements of Speaker 1. Figure 5.18 – 5.26 give 18 examples of such midsagittal configurations of the vocal tract estimated for Speaker 2. All three vowels /a/, /i/ and /u/ and both oral and nasalized vowel types were included in the examples to illustrate the corresponding articulatory features.

3.2 Area function

Figure 5.27 and 5.28 show the area functions computed based on the estimated midsagittal configurations of the vocal tract (e.g., Figure 5.9 – 5.26) and the adapted PONM for Speaker 1 and Speaker 2, respectively. The area function is represented by a 60-by-1 vector that contains the cross-sectional areas of 60 consecutive tubes from the glottis to the lip opening, where the x-axis corresponds to the index counted from the glottis and the y-axis corresponds to the area in the unit of mm^2. The nine subplots in both Figures 5.27 and 5.28 give the area functions of all vowel samples from the one of the following types: (a) oral /a/, (b) nasal /a/, (c) nasal /a/ with articulatory adjustment, (d) oral /i/, (e) nasal /i/, (f) nasal /i/ with articulatory adjustment, (g) oral /u/, (h) nasal /u/, and (i) nasal /u/ with articulatory adjustment.

3.3 Orthogonal articulatory modes

The first three orthogonal modes, taken together, account for over 95% variance of the nasal-oral area function difference (NA-O and N-O) for all vowels (/a/, /i/, /u/) and both speakers (Speaker 1 and Speaker 2). Figure 5.29 – 5.30 show the first three orthogonal modes for Speakers 1 and 2, respectively.

3.4 Acoustic features of the synthetic vowels

The lowest two formant frequencies F1 and F2 were computed for all synthetic vowel samples. Specifically, a Hamming window was applied to each vowel sample with its center at the midpoint of the vowel and a length of 512. Using the FFT function in MATLAB R2011_a and the FREQZ function in MATLABs Signal Processing Toolbox,16th-order Linear Predictive (LP) filters were designed for oral vowel samples and 30th-order filters were designed for nasal vowel samples. The LPC function detected formant peaks automatically in MATLAB.

In addition, figures of the 2048-point FFT spectrum with a LPC-based envelope were generated for each vowel sample. Figure 5.31 shows the spectra of nine vowel samples including an oral /a/, a nasal /a/, a nasal /a/ with articulatory adjustment, an oral /i/, a nasal /i/, a nasal /i/ with articulatory adjustment, an oral /u/, a nasal /u/ and a nasal /u/ with articulatory adjustment.

The spectra were also used to interactively check the LPC-derived formant frequencies. Whenever inconsistency happens between the LPC-derived formant frequency and the spectrum, the corresponding formant frequency was logged interactively in MATLAB by manually measuring the frequency of the harmonic closest to the formant peak in the spectrum. After such manual corrections, the formant frequencies F1 and F2 were plotted in Figure 5.32. According to ANOVA, the factor "vowel type" was found to have a significant effect on F1 for the following vowels: for Speaker 1, (1) /a/, $F(2, 166) = 31.737$, $p < .0001$, (2) /i/, $F(2, 264) = 36.16$, $p < .0001$; for Speaker 2, (1) /a/, $F(2, 178) = 14.153$, $< .0001$, (2) /i/, $F(2, 146) = 60.254$, $< .0001$, (3) /u/, $F(2, 84) = 15.127$, $p < .0001$. The following significant effects of "vowel type" were found on F2: for Speaker 1, (1) /a/, $F(2, 166) = 4.1479$, $p = .0175$, (2) /i/, $F(2, 264) = 163.15$, $p < .0001$, (3) /u/, $F(2, 127) = 9.7494$, $p = .0001$; for Speaker 2, (1) /i/, $F(2, 146) = 25.043$, $p < .0001$, (2) /u/, $F(2, 82) = 15.878$, $p < .0001$.

Furthermore, Tukey's post-hoc tests showed the following significant contrasts at the level of $< .05$ for Speaker 1: for /a/, (1) $F1_{NA} > F1_N$, $p < .0001$, (2) $F1_{NA} > F1_O$, $p < .0001$, (3) $F1_N < F1_O$, $p = .0006$, (4) $F2_{NA} < F2_O$, $p = .0358$; for /i/, (1) $F1_{NA} < F1_O$, $p < .0001$, (2) $F1_N < F1_O$, $p = 0.0031$, (3) $F2_{NA} < F2_N$, $p < .0001$, (4) $F2_{NA} < F2_O$, $p < .0001$, (5) $F2_N > F2_O$, $p < .0001$; for /u/, (1) $F2_{NA} < F2_N$, $p = 0.01$, (2) $F2_{NA} < F2_O$, $p < .0001$. The following significant contrasts were found for Speaker 2: for /a/, (1) $F1_{NA} > F1_N$, $p = .004$, (2) $F1_{NA} > F1_O$, $p < .0001$; for /i/, (1) $F1_{NA} < F1_N$, $p = 0.005$, (2) $F1_{NA} < F1_O$, $p < .0001$, (3) $F1_N < F1_O$, $p < .0001$, (4) $F2_{NA} < F2_N$, $p < .0001$, (5) $F2_{NA} > F2_O$, $p = .0039$, (6) $F2_N > F2_O$, $p < .0001$; for /u/, (1) $F1_{NA} < F1_O$, $p = .0023$, (2) $F1_N < F1_O$, $p < .0001$, (3) $F2_{NA} < F2_N$, $p < .0001$, (4) $F2_N > F2_O$, $p < .0001$.

3.5 Relationship between formant frequencies and articulatory adjustment

To examine the relationships between formant frequencies and articulatory adjustment, Figures 5.33–5.38 show the scatterplots of F1 and F2 versus the first three amplitude coefficients (pp1, pp2, pp3) of the corresponding orthogonal articulatory modes and their linear regression lines for each vowel and each speaker.

The following correlations (R) between formant frequency and amplitude coefficient were found to be significant for Speaker 1: for adjusted nasal /a/, (1) $R_{F1-pp1} = -0.3583$, $p = .0005$, (2) $R_{F1-pp2} = -0.3565$, $p = .0006$, (3) $R_{F1-pp3} = -0.2143$, $p = .0425$, (4) $R_{F2-pp1} = -0.4374$, $p < .0001$, (5) $R_{F2-pp3} = -0.2225$, $p = .0350$; for adjusted nasal /i/, (1) $R_{F1-pp3} = -0.3238$, $p = .0024$, (2) $R_{F2-pp1} = 0.8221$, $p < .0001$, (3) $R_{F2-pp3} = -0.8275$, $p < .0001$; for nasal /i/, (1) $R_{F2-pp1} = 0.4367$, $p = .0422$; for nasal /u/, (1) $R_{F2-pp1} = -0.4397$, $p = .0192$, (2) $R_{F2-pp2} = 0.4433$, $p = .0182$. For Speaker 2, the following correlations were significant: for adjusted /a/, (1) $R_{F1-pp1} = 0.5521$, $p < .0001$, (2) $R_{F1-pp2} = -0.5964$, $p < .0001$, (3) $R_{F1-pp3} = 0.5521$, $p < .0001$; for nasal /a/, (1) $R_{F1-pp3} = -0.5319$, $p = .009$, (2) $R_{F2-pp1} = 0.7035$, $p = .0002$, (3) $R_{F2-pp2} = -0.4256$, $p = .0429$; for nasal /i/, (1) $R_{F1-pp1} = -0.4470$, $p = .0482$, (2) $R_{F2-pp3} = -0.4575$, $p = .0425$; for nasal /u/, (1) $R_{F2-pp1} = 0.5990$, $p = .0053$, (2) $R_{F2-pp2} = 0.6322$, $p = .0028$.

3.6 Perceptual rating of nasality

3.6.1 Equalization of nasality scores

To reduce the effect of inter-listener variability of internal perceptual scale on the rating of nasality, the raw DME nasality scores were equalized using the psychophysical scaling method as follows (Engen, 1971):

1. Convert each response to its logarithm;
2. Calculate the mean of step 1, which is regarded as the logarithmic mean of each listener′s responses to all the stimuli;
3. Compute the mean of all the values obtained in step 2 to serve as the logarithmic grand mean of the responses from all listeners to all stimuli;
4. Subtract the logarithmic grand mean in step 3 from each of the individual mean log responses in step 2;
5. Subtract the value obtained in step 4 from the logarithmic responses made by each listener to each stimulus in step 1;

6. Convert each value obtained in step 6 to its exponential.

The values obtained in step 6 are regarded as the equalized nasality scores. In the following discussion, the terms such as "nasality score", "nasality rating" all refer to the equalized nasality score unless a special notification is given.

3.6.2 Comparison of nasality among O, N and NA

Figure 5.39 shows the barplots of the equalized nasality scores of O, N and NA for individual vowels /a/, /i/, /u/ as well as the average across three vowels. A linear mixed-effects (LME) model was fitted to examine the relationship between nasality and vowel type (O/N/NA) with vowel (/a/, /i/, /u/) as a random effect and listener as a nested random effect. The ANOVA on the linear mixed-effects model suggested a significant effect of vowel type on nasality ($F(3, 3038) = 268.0899$, $p < .0001$). Furthermore, a Tukey's post-hoc test suggested significant contrasts between NA and N as well as between NA and O. Specifically, the mean nasality score of NA is significantly lower than the mean nasality score of N ($mean[nasality_{NA} - nasality_N] = -95.130$, $p < .0001$) and is significantly higher than the mean nasality score of O ($mean[nasality_{NA} - nasality_O] = 27.638$, $p = .0001$).

The relationship between nasality and vowel type was also examined in individual vowels (/a/, /i/, /u/) by fitting a linear mixed-effects model on the relationship between nasality and vowel type (O/N/NA) for each vowel (/a/, /i/, /u/) with listener as a random effect. The ANOVA on the LME models for /a/, /i/ and /u/ found significant effects of vowel type on nasality for all vowels: (1) for /a/, $F(3, 1079) = 150.1022, p < .0001$; (2) for /i/, $F(3, 1244) = 579.6409, p < .0001$; (3) for /u/, $F(3, 711) = 158.6119, p < .0001$. Tukey's post-hoc tests on the LME models for individual vowels suggested the following significant contrasts between NA and the other two types of vowels (O, N): (1) for /a/, $mean[nasality_{NA} - nasality_N] = -145.99$, $p < .0001$; (2) for /i/, $mean[nasality_{NA} - nasality_N] = -54.620$, $p < .0001$; $mean[nasality_{NA} - nasality_O] = 34.411$, $p < .0001$; (3) for /u/, $mean[nasality_{NA} - nasality_N] = -94.15$, $p < .0001$; $mean[nasality_{NA} - nasality_O] = 38.18$, $p = 0.0179$.

3.7 Relationship between nasality and formant frequencies

To examine the relationship between nasality and acoustic features, Figure 5.40 shows the scatterplots of the equalized nasality scores versus the formant frequencies (F1, F2) for Speaker 1 with linear regression lines fitted to represent the nasality-acoustic relationships for O, N and NA, respectively. Similarly, Figure 5.41 shows the scatterplots and linear regression fits of the nasality-acoustic relationships for Speaker 2. As the

correlations between nasality and formant frequency are not prominent in Figures 5.40 and 5.41, the nasality scores were converted to their logarithmic values and correlated with the formant frequencies in Figures 5.42 and 5.43.

The following correlations between the logarithm of the equalized nasality scores and the formant frequencies for Speaker 1 were shown to be significant: for nasal /a/, $R_{F1-logNasality} = -0.5356$, $p = .0267$; for oral /a/, $R_{F2-logNasality} = 0.5408$, $p = .0009$; for adjusted nasal /u/, $R_{F1-logNasality} = -0.5741$, $p = .02$; for nasal /u/, $R_{F2-logNasality} = -0.6443$, $p = .0029$. For Speaker 2, the following correlations were found to be significant: for oral /a/, $R_{F1-logNasality} = 0.4229$, $p = .039$, $R_{F2-logNasality} = 0.6245$, $p = .0011$; for nasal /i/, $R_{F2-logNasality} = 0.4678$, $p = .018$.

3.8 Relationship between nasality and articulatory adjustment

To examine the relationship between nasality and articulatory adjustment, Figures 5.44–5.45 show the scatterplots of the equalized nasality scores versus the amplitude coefficients of the first three articulatory modes with the corresponding linear regression fits for Speakers 1 and 2, respectively. The nasality scores were also converted into a log scale and the logarithm of the nasality scores were correlated with the amplitude coefficients to result in the scatterplots and linear regression fits in Figures 5.46–5.47.

Examination on the correlations between the logarithm of the equalized nasality scores and the amplitude coefficients for Speaker 1 shows a significant correlation between $log(nasality)$ and pp1 for adjusted nasal /u/ ($R_{logNasality-pp1} = -0.5741$, $p = .02$). For Speaker 2, the correlation between $log(nasality)$ and pp2 is shown to be significant for nasal /i/ ($R_{logNasality-pp2} = 0.4678$, $p = .018$).

Chapter 4

Discussion

4.1 Oropharyngeal articulation of nasalized vowels

Based on the articulatory configuration estimated from the EMA measurements, the oropharyngeal articulations of oral and nasalized vowels can be compared. For Speaker 1, as shown in Figures 5.9–5.11, the tongue body is generally more anterior in nasalized /a/ compared to oral /a/. In addition, the tongue blade is higher in the nasalized /a/ from /ana/ compared to the oral /a/ from /ada/ (Figure 5.10); the entire tongue moves upward in the nasalized /a/ from /aŋa/ compared to the oral /a/ from /aga/ (Figure 5.11); the lip opening decreases in the nasalized vowels compared to their oral vowel counterparts (Figures 5.9 and 5.11). In Figures 5.12–5.14, the tongue blade moves upward and the lip opening size increases in the nasalized /i/ from /bim/ compared to the oral /i/ from /bip/ (Figure 5.12); the tongue tip moves posteriorly in the nasalized /i/ from /din/ compared to the oral /i/ from /dit/. In Figures 5.15–5.17, the tongue body is generally in a more posterior position with a smaller cross-sectional area at the velopharyngeal region in nasalized /u/ compared to oral /u/. For Speaker 2, the tongue body moves anteriorly in the nasalized /a/ from /pam/ compared to the oral /a/ from /pab/ (Figure 5.18) and in the nasalized /a/ from /kaŋ/ compared to the oral /a/ from /kak/ (Figure 5.20); in the nasalized /a/ from /tan/ (Figure 5.19), the tongue body moves more posteriorly with respect to the oral /a/ from /tad/ instead. For the vowel /i/, the tongue body moves posteriorly in the nasalized /i/ from /imi/ with respect to the oral /i/ from /ibi/ (Figure 5.21); the tongue body moves upward in the nasalized /i/ from /ini/ compared to the oral /i/ from /idi/ (Figure 5.22); the tongue tip moves anteriorly in the nasalized /i/ from /iŋai/ with respect to the oral /i/ from /igi/ (Figure 5.23). For the vowel /u/, the tongue blade generally moves downward in nasalized /u/ with respect to oral /u/ (Figures 5.24–5.26), which is sometimes accompanied by a posterior movement of the tongue back (Figures 5.24 and 5.26) or a downward and backward movement of the tongue tip (Figure 5.25).

These articulatory differences between oral and nasalized vowels are regarded as a results of nasalization, which accompany the lowering movement of the velum to influence the acoustic characteristics of nasalized vowels. Such oropharyngeal articulatory distinctions between oral and nasal(ized) vowels were found to exist

in a variety of languages. For the French nasal vowels /ã/ and /õ/, Zerling (1984) found the tongue body was retracted compared to their oral vowel counterparts; (Engwall et al., 2006) suggested French nasal vowels are in general produced with a larger front oral cavity with respect to the production of oral vowels; the tongue is lower and retracted in the French nasal vowel /ɛ̃/ versus the oral vowel /ɛ/; and tongue body is generally more anterior for nasal versus oral vowels in Hindi (Shosted et al., 2010); the tongue body is higher in the nasalized /i/ in English compared to its oral counterpart (Carignan et al., 2010). These findings show evidence in languages with and without phonemic nasal vowels that the oropharyngeal articulation of nasal(ized) vowels is not necessarily to be identical to the articulation of their oral vowel counterparts. In other words, the phonetic event of nasalization includes both velar lowering and oropharyngeal articulatory adjustment, the combination of which determines the shape of the vocal tract.

In this study, both overall patterns of articulatory adjustment and speech-dependent movements of individual articulators are found in the nasalized vowels to contrast with their oral vowel counterparts, resulting in a variability of oropharyngeal articulation. It indicates that, given the versatility of oropharyngeal articulation of nasalized vowels to interact with velar lowering, it is possible to adjust the oropharyngeal articulation in such a way that achieves the goal of compensating for the acoustic effect of velar lowering.

4.2 Acoustic changes as a result of articulatory adjustment

4.2.1 Spectral features

To examine the acoustic changes caused by the articulatory adjustment made by the model, Figure 5.31 shows the spectra of nine synthetic vowel samples including an oral /a/, a nasal /a/, a nasal /a/ with articulatory adjustment, an oral /i/, a nasal /i/, a nasal /i/ with articulatory adjustment, an oral /u/, a nasal /u/ and a nasal /u/ with articulatory adjustment.

First of all, the spectral characteristics of each O-N pair are compared to examine the primary acoustic outcome caused by VPO as a result of nasalization. As shown in Figure 5.31, the primary acoustic difference of each O-N pair lies in the low-frequency spectrum (below 3000 Hz) and, especially, the lowest two peaks above 200 Hz, which correspond to F1 and F2 (the lowest peak below 200 Hz corresponds to the fundamental frequency f0). For the low-back nasal vowel /a/, the amplitudes of both F1 and F2 are reduced and their bandwidths become wider as a result of nasal coupling. For the high-front nasal vowel /i/, a prominent extra peak between F1 and F2 is introduced as a result of the resonance of the nasal cavity; due to the extra nasal peak, the frequency of F2 increases. For the high-back nasal vowel /u/, the amplitude of F1 is reduced to some extent; the amplitude of F2 is largely attenuated and bandwidth of F2 becomes much

wider. In general, reduction of F1 amplitude and widening of F1 bandwidth are regarded as the primary effect of nasalization on F1, which is consistent with the finding of low-frequency spectral flattening by Maeda (1982a). For F2, the effect of nasalization is different on front versus back vowels. For the front vowel /i/, as F2 has a higher frequency than the nasal peak-zero pair (around 1500 Hz), the existence of the nasal zero shifts F2 upward to a higher frequency. For the back vowels /a/ and /u/, on the other hand, as F2 has a low frequency comparable to the frequency of the nasal resonance, the attenuation effect of the nasal zero causes a prominent reduction of the F2 intensity and a widening of the F2 bandwidth.

By comparing the spectra of each NA-N pair, the acoustic effect of articulatory adjustment can be distinguished. For the adjusted nasal vowel /a/, the amplitudes of both F1 and F2 are restored to be comparable to the amplitudes of F1 and F2 for the oral /a/; the bandwidths of F1 and F2 become narrower as a result of the increased amplitudes of F1 and F2. For the adjusted nasal /i/, the intensity of the nasal peak between F1 and F2 is largely attenuated and F2 is shifted downward to a lower frequency comparable to the F2 frequency of the oral /i/. For the adjusted nasal /u/, the amplitude of F2 is restored to a level comparable to the F2 amplitude of the oral /u/; as a result, the bandwidth of F2 becomes narrower compared to the nasal /u/.

These acoustic characteristics of NA are regarded as the effect of the speech-dependent articulatory adjustment generated by the speaker-adaptive model on the acoustics of nasal vowels. As shown in Figure 5.31, the articulatory adjustment successfully compensates for the acoustic effects caused by VPO, which include attenuation of F1 and F2 amplitudes (/a/, /u/), widening of F1 and F2 bandwidths (/u/), shifting of F2 frequency (/i/) and introduction of an extra peak-zero pair at around 1500 Hz caused by the resonance of the nasal cavity (/i/). As a result, the spectral look of NA is closer to the spectra of O compared to the large differences between the spectra of O and N.

Furthermore, by comparing the spectra of the adjusted nasal /i/ synthesized with the standard articulatory model in Figure 5.6 and the spectrum of the adjusted nasal vowels synthesized with the speaker-adaptive articulatory model in Figure 5.31, it is found that the articulatory adjustment generated by the speaker-adaptive model has a better effect on compensating for the acoustic outcome caused by VPO, which is evidenced by closer spectral looks of NA to the spectra of O in Figure 5.31. Such an effect is regarded as a result of the speaker adaptation of the articulatory model, which takes into account the inter-speaker variabilities of both PONM and articulatory space. Correspondingly, the articulatory adjustment generated by the speaker-adaptive model is customized with speaker-dependent features to give better acoustic compensation effects.

4.2.2 Formant frequencies

The lowest two formant frequencies F1 and F2 were compared among the three types of vowels (O/N/NA) to examine the effect of nasalization and articulatory adjustment on formant frequencies. Figure 5.32 shows the F1 and F2 formant frequencies of the synthetic vowels /a/, /i/ and /u/, grouped by vowel type (i.e., O, N, NA). By comparing the formant frequencies between N and O, the effect of nasalization on F1 and F2 includes a reduction of F1 frequency for nasal /a/ (Speaker 1), an increase of F2 frequency for nasal /a/ (Speaker 2), a reduction of F1 frequency for nasal /i/ (Speaker 1), an increase of F2 frequency for nasal /i/ (Speakers 1 and 2), a reduction of F1 frequency for nasal /u/ (Speaker 2) and an increase of F2 frequency for nasal /u/ (Speaker 2). Lowering of the F1 frequency for the low vowel /a/ is mostly likely to be attributed to the low-frequency resonances of the paranasal cavities, which tend to shift F1 to a low frequency (Pruthi et al., 2007). Shifting of F2 to a higher frequency for the back vowels /a/ and /u/ could be caused by the resonance of the nasal cavity, which usually has a frequency between 1000 Hz and 2000 Hz (Pruthi et al., 2007). On the other hand, the increase of F2 frequency for the front vowel /i/ is related to the existence of the nasal zero between F1 and F2, which shifts F2 to a higher frequency (Figure 5.31).

By comparing the formant frequencies between NA and N in Figure reffig:formants, the acoustic effect of articulatory adjustment can be examined. As a result of articulatory adjustment, it was found that 1) the F1 frequency of nasal /a/ is increased (Speakers 1 and 2), 2) the F2 frequency is reduced for nasal /i/ (Speakers 1 and 2), and 3) the F2 frequency of nasal /u/ is also reduced (Speakers 1 and 2). These changes compensate for the acoustic modification on formant frequency (F1, F2) caused by VPO.

Therefore, according to Figures 5.31 and 5.32, the articulatory adjustment introduces acoustic changes that compensate for the acoustic effect of VPO, which includes alteration of the oral formant structures and extra nasal peak-zeros caused by the resonances of the nasal and paranasal cavities, so that the formant structures (especially F1 and F2) are restored and the nasal resonant features are attenuated to result in a spectral look of the adjusted nasal vowel similar to the spectrum of the oral vowel.

4.3 Perceptual changes as a result of articulatory adjustment

According to the discussion above, the speaker-dependent articulatory adjustment generated by the speaker-adaptive model successfully compensates for the acoustic outcome caused by VPO. In other words, the acoustic goal of attenuating nasal resonant features and restoring oral formant structures is achieved through adjustment of articulation on individual bases. Perceptually, whether the articulatory adjustment achieves the auditory goal of reducing the perceived nasality is examined in Figure 5.39. As the nasality scores of all

nasal vowels (/a/, /i/, /u/) are significantly higher than the nasality scores of their oral vowel counterparts, the difference of nasality between these two types of vowels serves as the target to be reduced through articulatory adjustment. By comparing the nasality scores of NA and N, we found the articulatory adjustment significantly reduced the nasality of all three vowels (/a/, /i/, /u/) synthesized with the models of both speakers. Furthermore, the nasality of the adjusted nasal /a/ is not significantly different from the nasality of the oral /a/. These findings suggest the articulatory adjustment generated by the speaker-adaptive model successfully reduced the nasality rating, which is consistent with the acoustic compensation effect provided by the articulatory adjustment.

Specifically, shifting of F2 towards a higher frequency for the back nasal vowels /a/ and /u/ increases the intensity of the 1.6 kHz band of the spectrum to result in an increase of nasality compared to the oral vowels, according to Kataoka et al. (2001). On the other hand, the articulatory adjustments for nasal /a/ and /u/ are intended to shift F2 towards a lower frequency to decrease the intensity of the 1.6 kHz band, which results in a reduction of nasality of the adjusted nasal /a/ and /u/ with respect to their nasal vowel counterparts. For the front nasal vowel /i/, the prominent nasal peak caused by the resonance of the nasal cavity increases the intensity of the 1.6 kHz spectral band and, in turn, leads to increased nasality. The intensity of this nasal peak lying within the 1.6kHz band is attenuated after the articulatory adjustment (Figure 5.31), resulting in reduced nasality of the adjusted nasal /i/ with respect to the nasal /i/.

In addition to the overall reduction of the nasality of NA with respect to N, the variance of the nasality scores of NA is also reduced compared to N, suggesting better perceptual consistency of the nasal vowels after articulatory adjustment.

Therefore, the reduction of nasality as a result of articulatory adjustment is consistent with the acoustic changes of the adjusted nasal vowels with respect to their nasal vowel counterparts. Accordingly, the acoustic goal of attenuating nasal resonant features and restoring oral formant structures and the auditory goal of reducing the perception of nasality are both achieved through the articulatory adjustment generated by the speaker-adaptive model.

4.4 Articulatory adjustment that causes the acoustic and perceptual changes

To examine the specific articulatory adjustment generated by the model that causes the acoustic and perceptual changes as discussed above, the first three orthogonal articulatory modes of the synthetic nasal vowels with and without articulatory adjustment were derived and shown in Figures 5.29 and 5.30. As the artic-

ulatory modes were derived from the area function difference between the (adjusted) nasal vowel and oral vowel, the red solid curves and blue dashed curves in Figures 5.29 and 5.30 mark out the primary articulatory changes of NA and N versus O for Speakers 1 and 2, respectively.

By looking at the orthogonal modes of both N and NA for both speakers, it was found that in general, the first orthogonal mode that accounts for the largest variance of the nasal-oral area difference represents the overall articulatory adjustment pattern; the second orthogonal mode corresponds to a compensatory adjustment pattern that tunes the articulatory adjustment in the first mode; and the third orthogonal mode represents fine-tuning adjustments superimposed on the first orthogonal mode to shape the spectrum with more versatile acoustic features.

Specifically for Speaker 1, as shown in Figure 5.29, the first orthogonal mode of nasal /a/ indicates a construction inside the back and main oral cavity (AR2, AR23) with respect to oral /a/, which can be enabled by moving the tongue body upward toward the palate. The second orthogonal mode of nasal /a/ indicates a small enlargement in the posterior oral cavity (AR2) and a constriction in the front oral cavity (AR4) relative to oral /a/, which can be enabled by lowering the tongue back and moving the tongue tip upward. The third orthogonal mode of nasal /a/ represents a constriction in the posterior oral cavity (AR2) and some fine adjustments in the main and front oral cavities (AR23, AR4), which include a small enlargement in AR23, a constriction in AR4 and an enlargement of lip opening (AR5), compared to oral /a/. For the vowel /i/, the first orthogonal mode of nasal /i/ corresponds to a constriction inside the pharyngeal cavity (AR9) and an enlargement of lip opening relative to oral /i/, which can be enabled by moving the tongue backward and increasing the lip aperture. The second orthogonal mode of nasal /i/ stands for an enlargement in the lower pharynx (AR1), a constriction in the upper pharynx (AR9) and a construction at the lips with respect to oral /i/. The third orthogonal mode of nasal /i/ indicates a constriction in the pharyngeal cavity (AR1, AR9), a small enlargement in the back oral cavity (AR2) and a constriction at and behind the lips (AR5) relative to oral /i/, which can be enabled by moving the tongue body backward and downward and decreasing the lip opening. For the vowel /u/, the first orthogonal model of nasal /u/ indicates a small enlargement in the posterior oral cavity (AR2) and a small constriction in the main oral cavity (AR23) with respect to oral /u/, which can be realized by lowering the tongue back and moving the tongue blade upward. The second orthogonal mode of nasal /u/ corresponds to an enlargement in the pharynx (AR1, AR9), a constriction in the back oral cavity (AR2), an enlargement in the main oral cavity (AR23) and a constriction at the lips compared to oral /u/. Such a pattern can be achieved by moving the posterior tongue body forward and upward, lowering the tongue blade and decreasing the lip opening. The third orthogonal mode of nasal /u/ stands for an enlargement in the pharyngeal and posterior oral cavities (AR1, AR9, AR2), a small constriction in the main cavity (AR23) and an enlargement in the front oral

cavity (AR4) relative to oral /u/. These adjustments can be enabled by moving the posterior tongue body forward and downward, raising the tongue blade and lowering the tongue tip.

For the orthogonal modes of the adjusted nasal vowels in Figure 5.29, it is shown that the first orthogonal mode of the adjusted nasal /a/ includes a constriction in the posterior and main oral cavities (AR2, AR23), an enlargement in the front oral cavity (AR4) and a constriction at the lips compared to the oral /a/. Such adjustments can be enabled by raising the tongue body, lowering the tongue tip and decreasing the lip opening. The second orthogonal mode of the adjusted nasal /a/ indicates a decrease of the oral cavity volume relative to the oral /a/, which can be realized by lowering the tongue. The third orthogonal mode of the adjusted nasal /a/ indicates an enlargement of the posterior oral cavity (AR2), a constriction in the main oral cavity (AR23) and an enlargement in the front oral cavity and the lip opening (AR4, AR5) with respect to the oral /a/. These adjustments can be achieved by moving the posterior part of the tongue downward and the anterior part of the tongue upward, and increasing the lip opening. For /i/, the first orthogonal mode of the adjusted nasal /i/ corresponds to a constriction in the upper pharynx (AR9), enlargement of the main oral cavity (AR23) and the lip opening relative to the oral /i/. Such adjustments can be realized by moving the tongue body backward, lowering the tongue blade and increasing the lip opening. By comparing the first orthogonal modes of the adjusted nasal /i/ and the nasal /i/ (red solid and blue dashed curves in Figure 5.29, respectively), it was found that the two articulatory patterns are fairly similar. The second orthogonal model of the adjusted nasal /i/ is primarily an increase of lip opening with respect to the oral /i/. The third orthogonal mode of the adjusted nasal /i/ includes a constriction in the main oral cavity (AR23) and an enlargement of the front oral cavity (AR4) relative to the oral /i/, which can be enabled by moving the anterior tongue body upward and the tongue tip downward. For /u/, the first orthogonal mode of the adjusted nasal /u/ represents a small enlargement of the back oral cavity (AR2) and a larger enlargement of the main and front oral cavities (AR23, AR4) relative to the oral /u/, which can be achieved by lowering the tongue body, especially the tongue blade. The second orthogonal mode of the adjusted nasal /u/ stands for an overall enlargement of the pharynx and the back oral cavity (AR9, AR2), a constriction in the main and front oral cavities (AR23, AR4) and an increase lip opening. These adjustments can be achieved by moving the posterior tongue body forward and downward, raising the tongue blade and tip as well as increasing the lip aperture. The third orthogonal mode of the adjusted nasal /u/ corresponds to a small enlargement of the pharynx (AR1, AR9), a small constriction in the posterior oral cavity (AR2), a larger constriction in the front oral cavity (AR4) and a decreased lip opening relative to the oral /u/, which can be enabled by moving the posterior tongue body forward and upward, raising the anterior part of the tongue and decreasing the lip aperture.

For Speaker 2, according to Figure 5.30, the first orthogonal mode of nasal /a/ represents a constriction

in the oral cavity (AR2, AR23, AR4) with respect to the oral /a/, which can be achieved by raising the tongue towards the palate. The second orthogonal mode of nasal /a/ corresponds to an enlargement of the back oral cavity (AR2), a constriction in the main oral cavity (AR23), an enlargement of the front oral cavity (AR4) and a decreased lip opening compared to oral /a/. These adjustments can be realized by moving the posterior tongue body downward and the anterior tongue body upward, lowering the tongue tip and decreasing the lip opening. The third orthogonal mode of nasal /a/ includes an enlargement of the upper pharynx (AR9) and the back oral cavity (AR2), a constriction in the front oral cavity (AR4) and an increases lip opening with respect to oral /a/. By moving the tongue back forward and downward, raising the tongue tip and increasing the lip opening, the adjustments represented by the third mode can be achieved. For /i/, the first orthogonal mode of nasal /i/ corresponds to small enlargement of the pharynx (AR1, AR9) and the main oral cavity (AR23), a large enlargement of the front oral cavity (AR4) and a constriction at the lips compared to oral /i/. These adjustments can be enabled by moving the tongue body forward and downward, lowering the anterior part of the tongue and decreasing the lip opening. The second orthogonal mode of nasal /i/ is primarily an enlargement of the pharynx (AR1, AR9) compared to oral /i/, which can be realized by moving the tongue body anteriorly. The third orthogonal mode of the nasal /i/ includes a constriction in the main oral cavity (AR23) and an enlargement of the front oral cavity (AR4) relative to oral /i/, which can be achieved by raising the anterior tongue body and lowering the tongue tip. For /u/, the first orthogonal mode of nasal /u/ is primarily a constriction in the main oral cavity (AR23) relative to oral /u/, which can be realized by raising the anterior tongue body toward the palate. The second orthogonal mode of nasal /u/ suggests an enlargement of the pharynx (AR1, AR9) and a constriction in the oral cavity (AR2, AR23, AR4) compared to oral /u/. Such adjustments could be achieved by moving the tongue body forward and upward. The third orthogonal mode of nasal /u/ indicates a small enlargement in the upper pharynx (AR9), a larger enlargement in the main oral cavity (AR23) and a small constriction in the front oral cavity (AR4) compared to oral /u/. These adjustments can be enable by moving the tongue body forward and downward and raising the tongue tip toward the palate.

For the orthogonal modes of the adjusted nasal vowels of Speaker 2 (Figure 5.30), the first orthogonal mode of the adjusted nasal /a/ corresponds to an enlargement of the oral cavity (AR2, AR23, especially AR4) with respect to the oral /a/, which can be realized by lowering the tongue body. The second orthogonal mode of the adjusted nasal /a/ indicates an enlargement of the main oral cavity (AR2, AR23), a constriction in the front oral cavity (AR4) and an increase lip opening compared to the oral /a/. These adjustments can be enabled by lowering the tongue body, raising the tongue tip and increasing lip opening. The third orthogonal mode of the adjusted nasal /a/ includes an enlargement of the back oral cavity (AR2), a constriction in the main oral cavity (AR23) and an enlargement of the front oral cavity (AR4) and lip opening compared to

the oral /a/. Such adjustments can be realized by lowering the posterior part of the tongue body and raising the anterior part of the tongue body and increasing lip opening. For /i/, the first orthogonal mode of the adjusted nasal /i/ corresponds to enlargement of the upper pharynx (AR9) and the front oral cavity (AR4) relative to the oral /i/, which can be enabled by moving the tongue body forward and the tongue tip downward. The second orthogonal mode of the adjusted nasal /i/ represents a small constriction in the upper pharynx (AR9), a small enlargement of the main oral cavity (AR23) and a large increase of lip opening compared to the oral /i/, where these adjustments can be realized by moving the tongue body backward, lowering the tongue tip and increasing the lip aperture. The third orthogonal mode of the adjusted nasal /i/ indicates a constriction in the pharynx (AR1, AR9) and an enlargement of the front oral cavity (AR4) with respect to the oral /i/, which can be achieved by moving the tongue body backward and lowering the tongue tip. For /u/, the first orthogonal mode of the adjusted nasal /u/ represents an overall increase of volume in the upper pharynx and the oral cavity (AR9, AR2, AR23) relative to the oral /u/, which can be enabled by moving the tongue body forward and downward. The second orthogonal mode of the adjusted nasal /u/ indicates a constriction in the upper pharynx (AR9) and an enlargement of the oral cavity (AR2, AR23) compared to the oral /u/, where such adjustments could be realized by moving the tongue body backward and downward. The third orthogonal mode of the adjusted nasal /u/ corresponds to an enlargement of the back oral cavity (AR2) and a constriction in the front oral cavity (AR23) with respect to the oral /u/, where the adjustments can be achieved by lowering the posterior tongue body and raising the tongue blade.

By comparing the orthogonal modes of NA and N (i.e., red solid and blue dashed curves in Figures 5.29 and 5.30, respectively), the principal patterns of articulatory adjustment generated by the models of Speaker 1 and Speaker 2 can be visualized straightforwardly.

4.5 Relationship among articulatory adjustment, formant frequency and nasality

4.5.1 Relationship between articulatory adjustment and formant frequency

To explore the relationship between the articulatory adjustment patterns indicated by the orthogonal modes and the formant frequencies (F1, F2), Figures 5.33–5.38 show the scatterplots of the formant frequencies versus the amplitude coefficients of the first three orthogonal modes for each vowel and each speaker. The slope of the correlation between formant frequency and amplitude coefficient indicates the degree of the variance of formant frequency accounted for by each orthogonal mode. A positive correlation indicates the articulatory adjustment represented by the orthogonal mode contributes to the increase of the corresponding

formant frequency, whereas a negative correlation suggests the articulatory adjustment represented by the orthogonal mode contributes to the decrease of the corresponding formant frequency.

Based on the difference of formant frequency between N and O for Speaker 1 in Figure 5.32, the articulatory gestures represented by the first orthogonal mode contribute to the lowering of F1 frequency of nasal /a/ with respect to oral /a/ (Figure 5.33(a)). The articulatory adjustments in the first and third orthogonal modes contribute to a decreased F1 frequency of nasal /i/ relative to oral /i/ (Figure 5.34(a),(c)). In addition, the first orthogonal mode also contributes to the raising of F2 frequency of nasal /i/ with respect to oral /i/ (Figure 5.34(d)). For nasal /u/, the articulatory adjustment represented by the first orthogonal mode leads to a reduction of F1 frequency compared to oral /u/ (Figure 5.35(a)). For Speaker 2, the articulatory adjustments represented by the first and third modes contribute to the raising of F2 frequency of nasal /a/ with respect to oral /a/ (Figure 5.36(d),(f)). The first and third orthogonal modes are responsible for the lowering of F1 frequency of nasal /i/ relative to oral /i/ (Figure 5.37(a),(c)), whereas the first and second orthogonal modes contribute to the raising of F2 frequency of nasal /i/ with respect to oral /i/ (Figure 5.37(d),(e)). For nasal /u/, the first and second orthogonal modes are responsible for both the lowering of F1 frequency and the raising of F2 frequency compared to oral /u/ (Figure 5.38(a),(b),(d),(e)).

According to the difference of formant frequency between NA and O in Figure 5.32, the relationship between each orthogonal mode of NA and the NA-O formant frequency difference was examined. For Speaker 1, the articulatory adjustments in the first and third orthogonal modes contribute to the lowering of F2 frequency of the adjusted nasal /a/ with respect to the oral /a/ (Figure 5.33(d),(f)). For the adjusted nasal /i/, all of the three orthogonal modes contribute to the lowering of F1 frequency (Figure 5.34(a)–(c)) and the second and third orthogonal modes are responsible for the lowering of F2 frequency with respect to the oral /i/ (Figure 5.34(e),(f)). For the adjusted nasal /u/, the first and third orthogonal modes contribute to the lowering of F2 frequency relative to the oral /u/ (Figure 5.35(d),(f)). For Speaker 2, the first and third orthogonal modes contribute to the raising of F1 frequency of the adjusted nasal /a/ relative to the oral /a/ (Figure 5.36(a),(c)). For the adjusted nasal /i/, the first and third orthogonal modes are responsible for the lowering of F1 frequency compared to the oral /i/ (Figure 5.37(a),(c)). For the adjusted nasal /u/, the articulatory gestures represented by the first and second orthogonal modes contribute to the lowering of F1 frequency with respect to the oral /u/ (Figure 5.38(a),(b)).

4.5.2 Relationship between nasality and formant frequency

The relationships between formant frequencies (F1, F2) and nasality are shown in Figures 5.40–5.43. By examining the slope of the correlations, the effect of formant frequencies on nasality rating can be evaluated.

For Speaker 1, the reduction of F1 frequency for nasal /a/ relative to oral /a/ corresponds to increased nasality (Figure 5.42(a)). Increased F2 frequency for nasal /i/ with respect to oral /i/ is related to increased nasality (Figure 5.42(e)). Decreased F2 frequency for nasal /u/ relative to oral /u/ is correlated with increased nasality (Figure 5.42(f)). For Speaker 2, increase F2 frequency for nasal /a/ relative to oral /a/ corresponds to increased nasality (Figure 5.43(d)). A decrease of the F1 frequency and an increase of the F2 frequency for nasal /i/ with respect to oral /i/ are related to increased nasality and decreased nasality, respectively (Figure 5.43(b),(e)). For /u/, a decrease of F1 frequency and an increase of F2 frequency for nasal /u/ relative to oral /u/ are correlated with decreased and increased nasality, respectively (Figure 5.43(c),(f)).

Based on the correlations between nasality and formant frequencies, it was found that decreased F1 frequency for nasal /a/ (Speaker 1), raised F2 frequency for nasal /a/ (Speaker 2), increased F2 frequency for nasal /i/ (Speakers 1 and 2), and decreased F1 frequency for nasal /u/ (Speakers 1 and 2) served as the acoustic correlates of nasality.

With regard to the adjusted nasal vowels, for Speaker 1, increased F1 frequency and decreased F2 frequency for the adjusted nasal /a/ relative to oral /a/ both correspond to a reduction of nasality (Figure 5.42(a),(d)). Decreased F1 frequency for the adjusted nasal /i/ with respect to oral /i/ is related to reduced nasality (Figure 5.42(b)). For Speaker 2, increased F1 frequency for the adjusted nasal /a/ relative to oral /a/ is correlated with reduced nasality (Figure 5.43(a)). Decreased F1 frequency and increased F2 frequency for the adjusted nasal /i/ relative to oral /i/ correspond to increased nasality and reduced nasality, respectively (Figure 5.43(b),(e)).

According to the nasality-formant frequency correlations for the adjusted nasal vowels, increased F1 frequency for the adjusted nasal /a/ (Speakers 1 and 2), decreased F2 frequency for the adjusted nasal /a/ (Speaker 1), reduced F1 frequency for the adjusted nasal /i/ (Speaker 1), and increased F2 frequency for the adjusted nasal /i/ (Speaker 2) are regarded as the acoustic cues related to the reduction of nasality.

4.5.3 Relationship between articulatory adjustment and nasality

To examine how articulatory adjustment affects the perception of nasality, Figures 5.44–5.47 show the relationship between nasality and the amplitude coefficients of the first three orthogonal modes.

For Speaker 1, the positive correlation between nasality and the first amplitude coefficient for nasal /a/ in Figure 5.46(a) suggests the articulatory gestures in the first orthogonal mode contributes to increased nasality of nasal /a/. The positive correlation between nasality and the first amplitude coefficient for nasal /i/ in Figure 5.46(d) indicates the articulatory pattern represented by the first orthogonal mode is correlated with increased nasality of the nasal /i/. For Speaker 2, the positive correlation between nasality and the second

amplitude coefficient for nasal /a/ in Figure 5.47(b) indicates a direct relationship between the articulatory pattern in the second orthogonal mode and increase of nasality for nasal /a/. For nasal /i/, the positive correlation between nasality and the second amplitude coefficient in Figure 5.47(e) suggests the articulatory gestures in the second orthogonal mode are responsible for the raising of nasality for the nasal /i/.

With regard to the adjusted nasal vowels for Speaker 1, the negative correlation between nasality and the first amplitude coefficient for the adjusted nasal /a/ in Figure 5.46(a) suggests the articulatory adjustment represented by the first orthogonal mode is responsible for the reduction of nasality for the adjusted nasal /a/. For the adjusted nasal /i/, the negative correlations between nasality and the second and third amplitude coefficients in Figure 5.46(e)–(f) indicate the articulatory adjustment patterns in the second and third orthogonal modes both contribute to the reduction of nasality. For the adjusted nasal /u/, the negative correlation between nasality and the first amplitude coefficient in Figure 5.46(g) suggests a direct relationship between the articulatory adjustment in first orthogonal mode of the adjusted nasal /u/ and reduction of nasality. For Speaker 2, the negative correlations between nasality and the second and third amplitude coefficients for the adjusted nasal /u/ in Figure 5.47(h)–(i) suggest the articulatory adjustment patterns in the second and third orthogonal modes contribute to the reduction of nasality for the adjusted nasal /u/.

4.5.4 The effect of articulatory adjustment on the acoustics and nasality of nasal vowels

Based on the relationship among formant frequency, nasality and articulatory pattern as discussed above, the acoustic correlates of nasality for Speaker 1 have been found as (1) a reduction of F1 frequency for /a/, which is related to an upward movement of the tongue body (especially the tongue blade) (Figure 5.29(a)), (2) an increase of F2 frequency for /i/, which is related to a backward and downward movement of the tongue and increased lip opening (Figure 5.29(d)), (3) a reduction of F2 frequency for /u/, which is related to a small forward and upward movement of the tongue (Figure 5.29(g)). The acoustic correlates of nasality for Speaker 2 include (1) an increase of F2 frequency for /i/, which is related to a forward movement of the tongue body (Figure 5.30(e)), (2) a reduction of F1 frequency for /u/, which is related to an upward movement of the tongue body (Figure 5.30(g)).

To compensate for the acoustic cues related to increased nasality for Speaker 1, (1) the tongue body should move upward and the tongue tip should move downward with decreased lip opening to generate the articulatory pattern of the adjusted nasal /a/ in Figure 5.29(a), which is related to a decrease of F2 frequency and a reduction of nasality; (2) the anterior tongue body should move upward with the tongue tip pointing downward to generate the articulatory pattern of the adjusted nasal /i/ in Figure 5.29(f), which corresponds

to reductions of both F1 and F2 frequencies and a reduction of nasality. For Speaker 2, the tongue body should move backward and the tongue tip should move downward to generate the articulatory pattern of the adjusted nasal /i/ in Figure 5.30(f), which leads to an increase of F2 frequency and a reduction of nasality.

By comparing the results of Speaker 1 and Speaker 2, it is found that different speakers share common acoustic cues of nasality such as reduced F1 frequency for nasal /a/ and increased F2 frequency for nasal /i/, and speaker-dependent acoustic features related to nasality such as reduced F2 frequency for nasal /u/ by Speaker 1 and reduced F1 frequency for nasal /u/ by Speaker 2. On the other hand, these common or speaker-dependent acoustic correlates of nasality are produced with different articulatory patterns (Figures 5.29 and 5.30), which require speaker-dependent articulatory adjustments to reduce the perception of nasality. Such speaker-dependent articulatory adjustments shown in Figures 5.29 and 5.30 have different degrees of correlation with nasality (Figures 5.46 and 5.47). Some articulatory adjustments such as the pattern indicated by the third orthogonal mode of the adjusted nasal /i/ for Speaker 1 have consistent effects on compensating for the shifting of formant frequency caused by nasalization and reducing the perception of nasality. Other articulatory adjustments such as the first orthogonal mode of the adjusted nasal /a/ that result in a reduction of nasality have a different effect on the nasal acoustic cues, which enhances the distinction of F1 frequency and reduces the distinction of F2 frequency between nasal and oral vowels. In such cases, it is most likely that additional acoustic cues other than F1 and F2 formant frequencies play a role in determining nasality. Especially for the vowel /u/, as shown in Figures 5.46 and 5.47, the correlation between nasality and the articulatory adjustment patterns for the adjusted nasal /a/ is either not prominent or in a different direction from expected, which provides indirect evidence of additional acoustic cues of nasality other than F1 and F2. Actually, the importance of the overall spectral shape in determining nasality has been found in previous studies Ito et al. (2001); Kataoka et al. (2001). Future studies may focus on finding an acoustic target that takes into account the overall spectral shape without causing an unreasonable increase of computational load.

To summarize, the articulatory adjustments generated by the speaker-adaptive model are speaker- and speech-dependent with a common goal of attenuating the nasal acoustic cues and reducing the perception of nasality. According to the definition of motor equivalence (Hughes and Abbs, 1976; Maeda, 1990), the deviation from the acoustic/auditory goal as a result of the shifted position of an articulator (in this case, lowering of the velum) can be compensated for by adjusting and coordinating other articulators so that the acoustic/auditory goal defined by an oral vowel target can be achieved. The articulatory adjustment generated by the speaker-adaptive model, although has speaker- and speech-dependent features, share some common patterns as indicated by the principal orthogonal modes derived from the difference of area functions of NA and O. These articulatory patterns contain gross adjustment patterns (e.g., 1^{st} orthogonal mode),

regional compensation patterns (e.g., 2^{nd} orthogonal mode) and fine-tuning adjustment patterns (e.g., 3^{rd} orthogonal mode). For vowels such as nasal /a/, the low-frequency spectrum has a relatively simple structure (Figure 5.31), so a gross articulatory adjustment pattern manages to achieve the goal of compensating for the acoustic features caused by VPO. On the other hand, for vowels such as nasal /i/ with a more complex low-frequency spectral structure, it relies on more refined adjustments to tune the acoustic changes caused by the gross articulatory adjustment.

4.6 Clinical implications

The approach that uses computer simulation of speaker-dependent articulatory adjustment to reduce nasality has some clinical implications. First, it provides a way of articulatory learning through human-computer interaction. Given an acoustic target, the speaker-adaptive articulatory model is able to adjust the articulatory configuration of the vocal tract to compensate for the acoustic outcome caused by VPO without moving the velum. The resultant articulatory configuration can be applied as a learning target to train patients with VPI to change their articulation accordingly.

A more straightforward implication is to help clinicians to determine what articulatory gestures should be made or avoided to reduce the perception of nasality. Individuals with VPI sometimes spontaneously develop maladaptive articulatory patterns such as pushing the tongue dorsum up against the velum in an attempt to achieve velopharyngeal closure. According to the highly-variable articulatory patterns indicated by the orthogonal modes, such an over-simplified articulatory behavior could be maladaptive and should be identified by clinicians and modified therapeutically. Based on Figures 5.29 and 5.30, a common articulatory strategy is to adjust the tongue placement in a specific direction to shape the spectrum with desired acoustic features. To further refine the spectral features, fine-tuning adjustments such as the lip movement are included in accordance to the acoustic outcome of the tongue movement so that the acoustic effects of the tongue adjustment can be tuned.

Work in this area could be extended to include additional vowels and connected speech. Articulatory adjustment strategies could be different for vowels other than /a/, /i/, /u/ because the acoustic cues for nasality may be slightly different for those vowels. In connected speech, it would be important to first identify how perception of nasality plays out for vowels versus consonants and whether articulatory adjustments would be effective in changing the percept.

Chapter 5

CONCLUSION

With a long-term goal of developing a therapeutic tool to assist individual-based diagnosis and treatment of hypernasal speech, this study constructed a speaker-adaptive articulatory model to simulate articulatory adjustments to compensate for the acoustic outcome caused by excessive velopharyngeal opening and in turn, to reduce the perception of nasality. The construction of the speaker-adaptive model follows a three-step paradigm: (1) fitting point-wise articulatory positions from a database measured by EMA to the framework of Childers's standard vocal tract model; (2) adjusting the PONM of the model to minimize the acoustic discrepancy between the model simulation and the acoustic target using the simulated annealing algorithm; (3) adjusting the articulatory space of the model to fit individual articulatory features by adapting the movement ranges of all articulators.

With the speaker-adaptive articulatory model, the articulatory configurations of the oral and nasal vowels in the database were simulated and synthesized. Given the acoustic targets derived from the oral vowels in the database, speech-dependent articulatory adjustments were generated and simulated to compensate for the acoustic deviation from the target caused by VPO. The corresponding nasal vowels with model-generated articulatory adjustment were synthesized and, taken together with the synthetic oral and nasal vowels as generated above, served as the perceptual stimuli for a listening task of nasality rating.

Comparison of the acoustic features among the oral vowels, nasal vowels and nasal vowels with articulatory adjustment suggests (1) the altered formant structures due to nasal coupling, including shifted formant frequency, attenuated formant intensity and expanded formant bandwidth, are restored and (2) the extra peaks and zeros caused by the resonances of the nasal and paranasal cavities are attenuated by the articulatory adjustment generated by the speaker-adaptive model. Furthermore, the speaker-dependent articulatory adjustment generated by the speaker-adaptive model provides better compensation effects for acoustic outcome of VPO compared to the articulatory adjustment generated by a standard articulatory model in the pilot study. It suggests that, by customizing the articulatory model with individual-based anatomical and articulatory features, it can generate more efficient articulatory adjustment to compensate for the acoustic outcome caused by VPO. In addition to the acoustic effects, the articulatory adjustment generated by the

speaker-adaptive model also manages to significantly reduce the perceived nasality for all three vowels /a/, /i/ and /u/, suggesting achievement of the auditory goal of compensating for the perceptual deviation (in this case, hypernasality) caused by VPO.

To examine the specific articulatory adjustment made by the model, a set of orthogonal modes were decomposed from the difference of area functions between the adjusted nasal and oral vowels. Both gross articulatory adjustment patterns and fine-tuning adjustments were found in the principal orthogonal modes, which were taken together to shape the spectrum of the vowel with desired acoustic features. For /a/ and /i/, direct relationships were found among the acoustic features, nasality, and the principal articulatory adjustment patterns. Specifically, the articulatory adjustments indicated by the principal orthogonal modes of the adjusted nasal /a/ and /i/ were directly correlated with the attenuation of the acoustic cues of nasality (i.e., shifting of F1 and F2 frequencies) and the reduction of nasality rating. For /u/, such a direct relationship among the acoustic features, nasality and articulatory adjustment was not found, suggesting the possibility of additional acoustic cues other than F1 and F2 to account for the perception of nasality.

To summarize, this study demonstrates the feasibility of using articulatory adjustment to compensate for the acoustic outcome caused by VPO and in turn, to reduce nasality through model simulation. By customizing the articulatory model with speaker-dependent anatomical and articulatory features, it is able to simulate individual-based articulatory adjustments that can be applied in clinical settings as the articulatory target to correct the maladaptive articulatory behaviors developed spontaneously by speakers with hypernasal speech. Such a model simulation provides an intuitive way of articulatory learning and self-training through model-speaker interaction, which may have potential clinical applications in development of an individualized treatment of hypernasal speech.

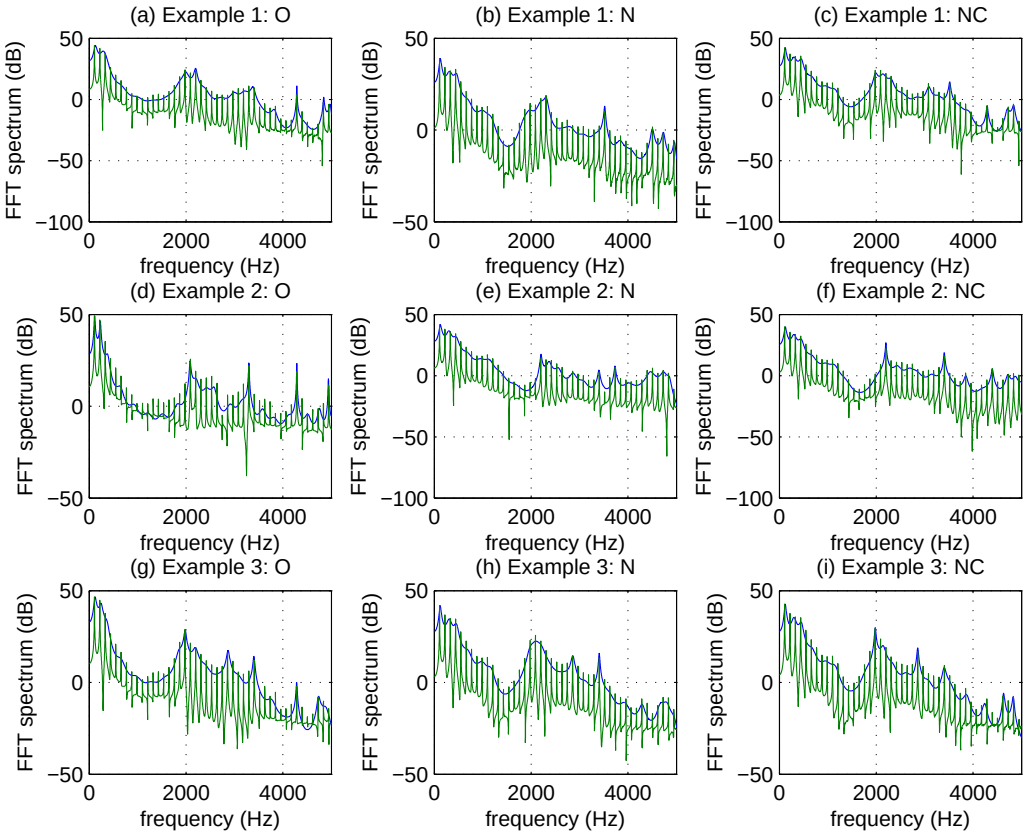


Figure 5.1: Spectra of three synthetic /i/ vowel sets, where (a)-(c) correspond to the first set of oral vowel, nasal vowel without adjustment and nasal vowel with adjustment. (d)-(f), (g)-(i) correspond to the second and third vowel sets, respectively.

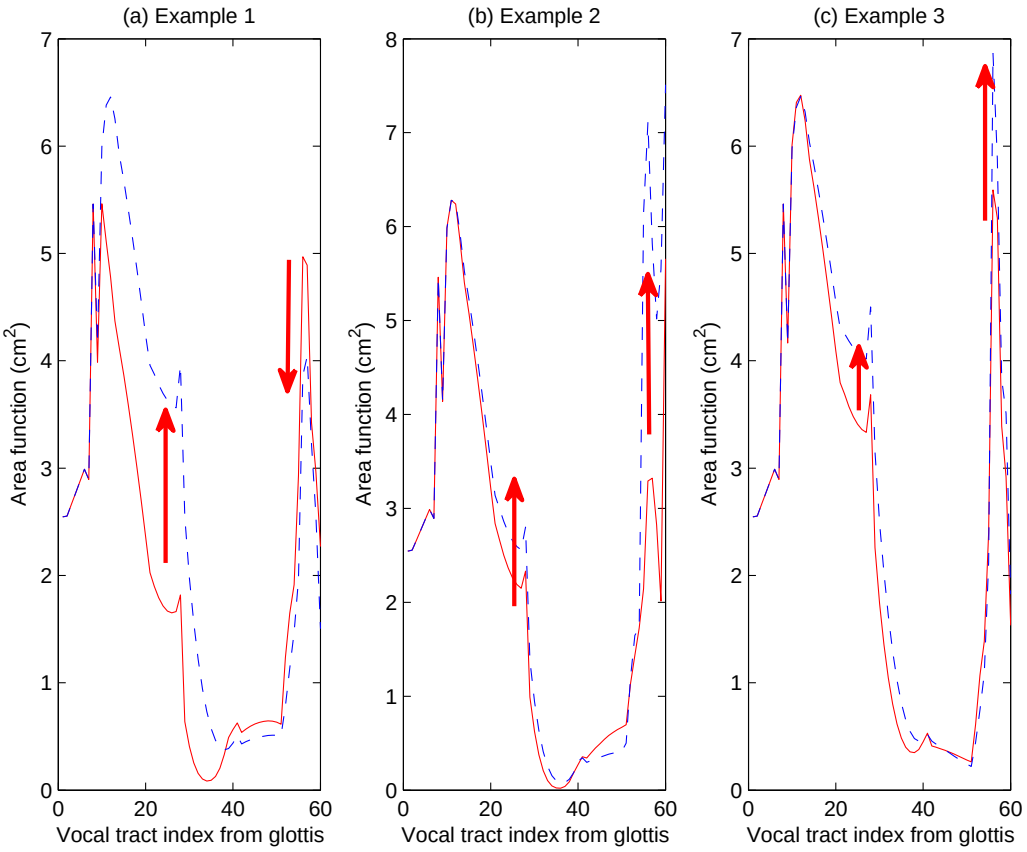


Figure 5.2: (a) Vocal tract area functions for a nasal vowel without adjustment (solid line, corresponding to the spectrum in Figure 5.6 (b)) and a nasal vowel with adjustment (dashed line, corresponding to the spectrum in Figure 5.6 (c)); (b), (c) Vocal tract area functions for the nasal vowels with/without adjustment in the second (corresponding to Figure 5.6 (e), (f)) and third (corresponding to Figure 5.6 (h), (i)) vowel sets, respectively. The arrows mark the critical changes of vocal tract shape after articulatory adjustment.

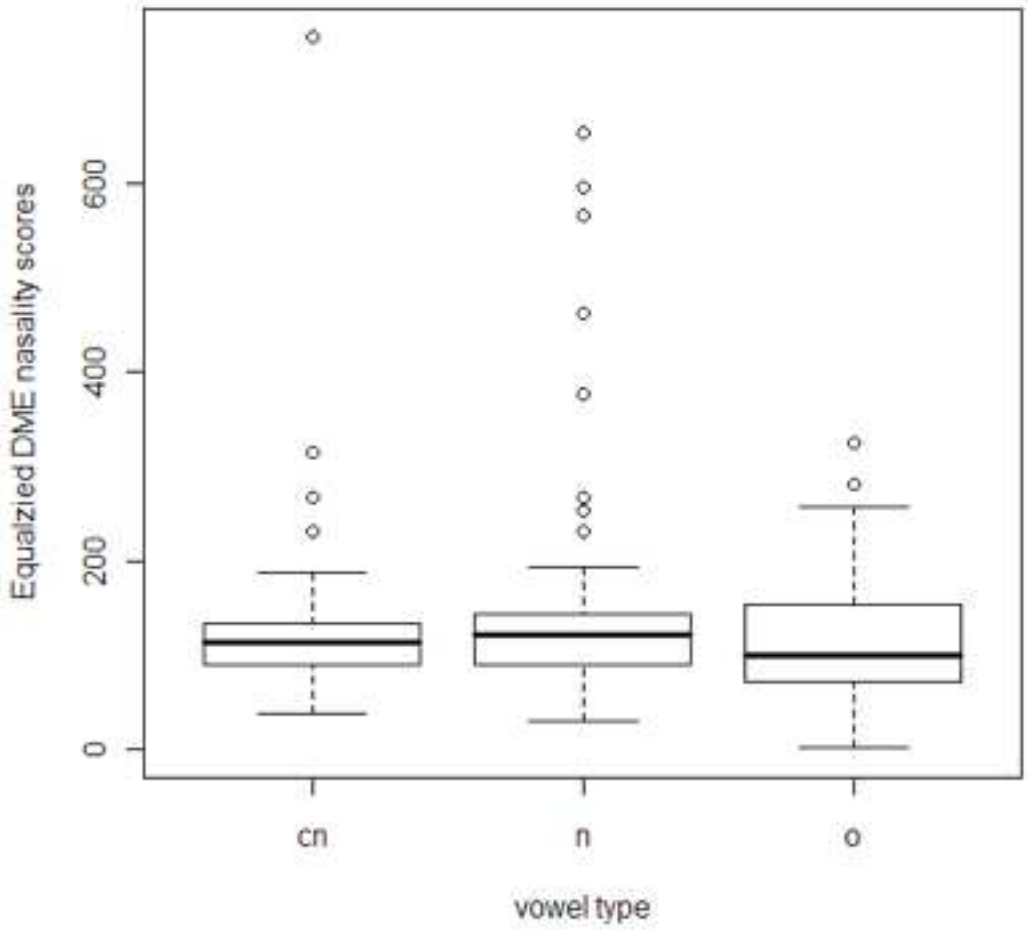


Figure 5.3: Box plots for equalized DME nasality scores grouped by vowel type, where "cn," "n," "o" stand for nasal vowels with and without adjustment, and oral vowels, respectively.

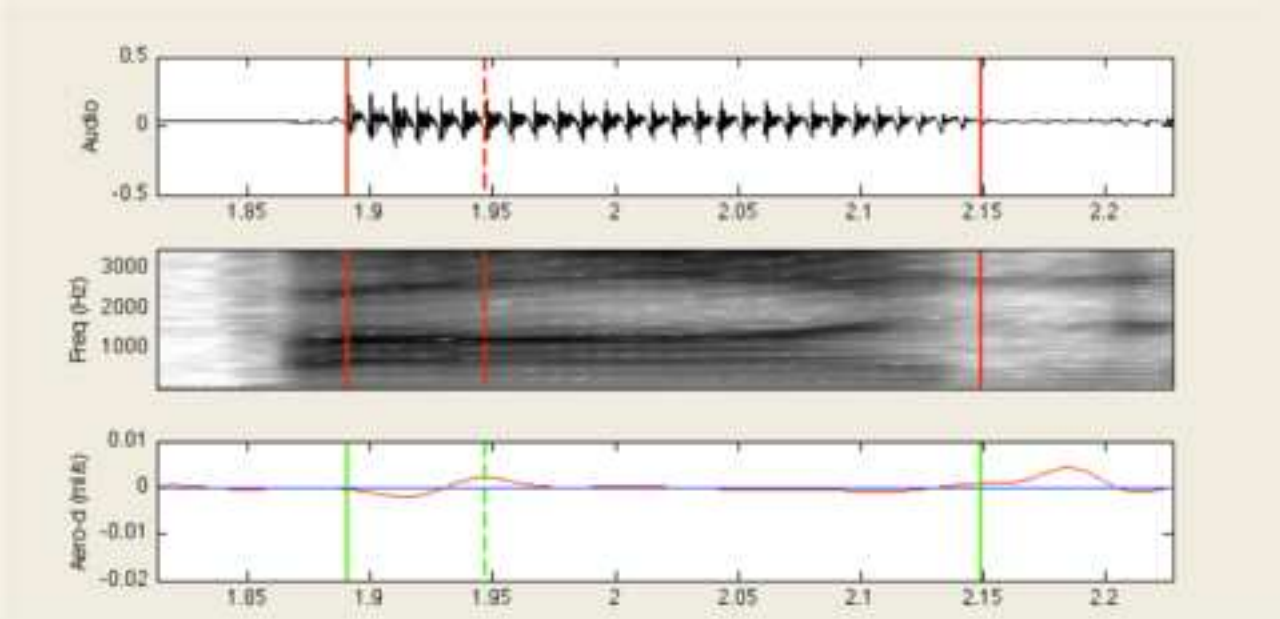


Figure 5.4: Acoustic and aerodynamic annotation scheme of /baŋ/. The three panels from top to bottom correspond to the acoustic waveform, spectrogram and nasal airflow velocity, respectively. The two vertical solid lines mark the onset and offset of the vowel /a/, while the vertical dashed line marks the onset of anticipatory nasalization. The onset and offset of vowel are determined based on the waveform and the onset of nasalization is determined by the first velocity peak after the vowel onset.

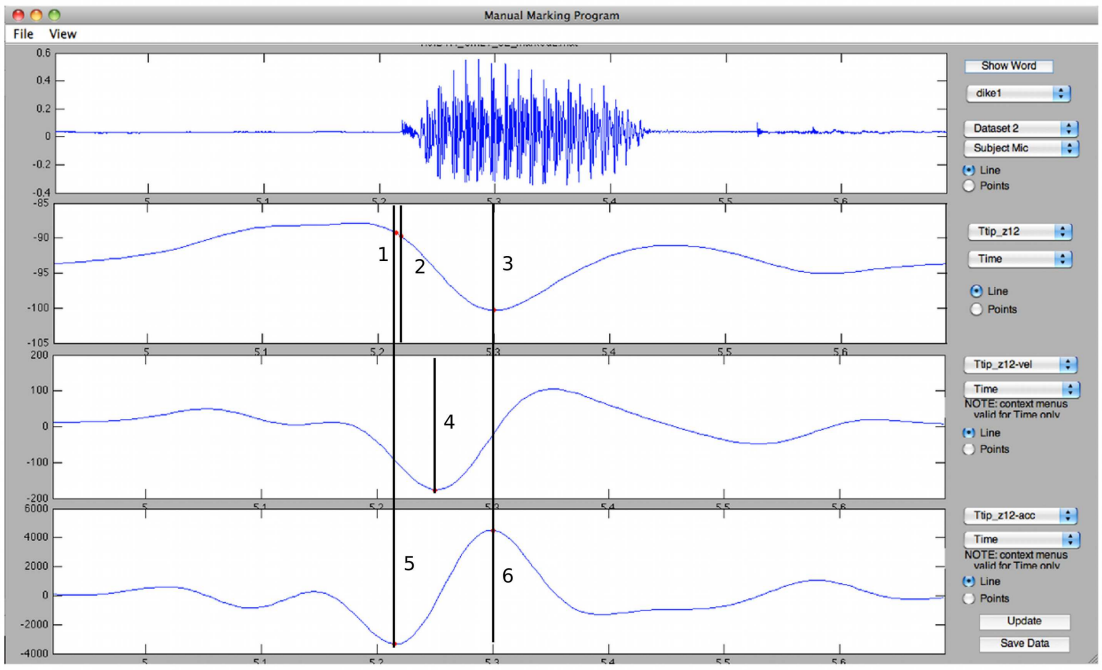


Figure 5.5: Articulatory annotation scheme of the word 'dike'. The four panels from top to bottom correspond to the acoustic waveform, tongue tip displacement, velocity and acceleration, respectively. Lines 1 and 3 mark the beginning and end of the opening phase of the initial consonant /d/, respectively, which are determined by the maximum acceleration of the lowering movement of tongue tip (marked by line 5) and the maximum deceleration of the tongue tip lowering (marked by line 6).

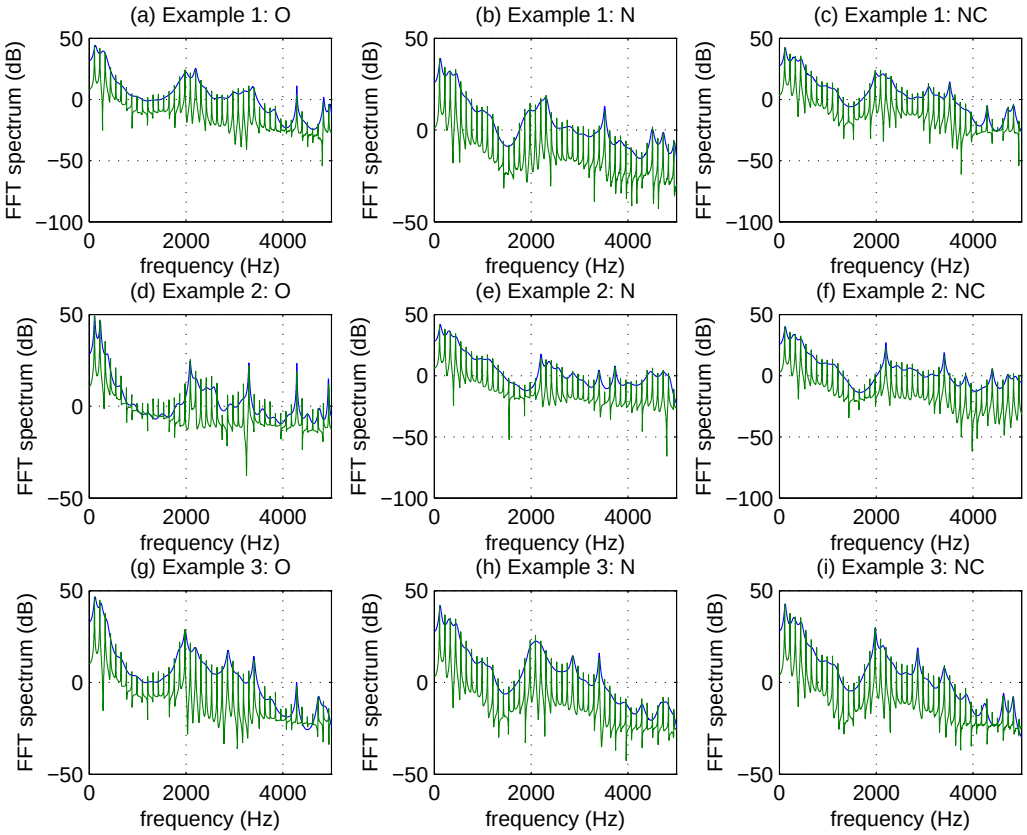


Figure 5.6: Spectra of three synthetic /i/ vowel sets, where (a)-(c) correspond to the first set of oral vowel, nasal vowel without adjustment and nasal vowel with adjustment. (d)-(f), (g)-(i) correspond to the second and third vowel sets, respectively.

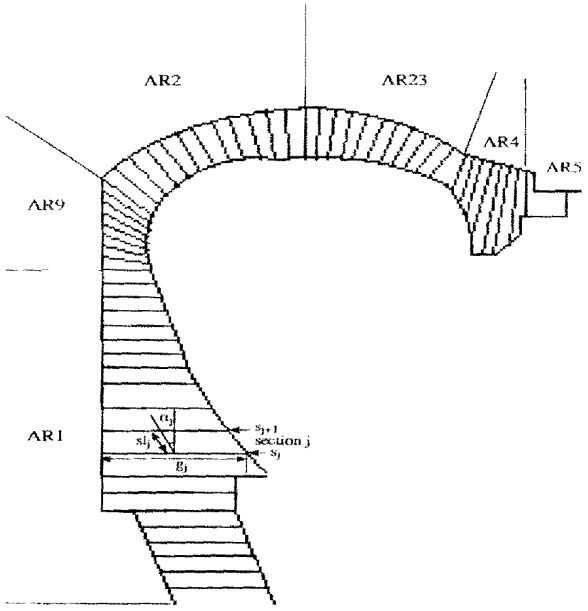


Figure 5.7: Graphical representation of Childer's vocal tract model. The entire vocal tract was divided into 60 consecutive tubular sections, the areas of which compose the area function of the vocal tract. According to the anatomy of the vocal tract, the 60 tubes were further grouped into six sections, namely, AR1, AR9, AR2, AR23, AR4 and AR5, from the glottis to the lips.

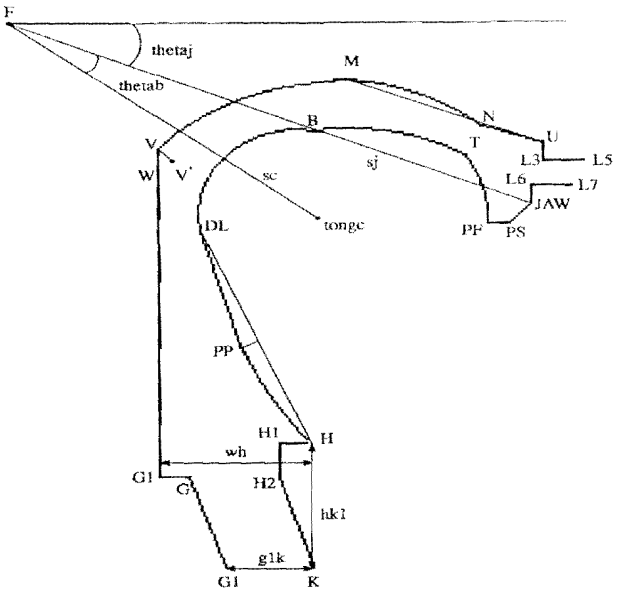
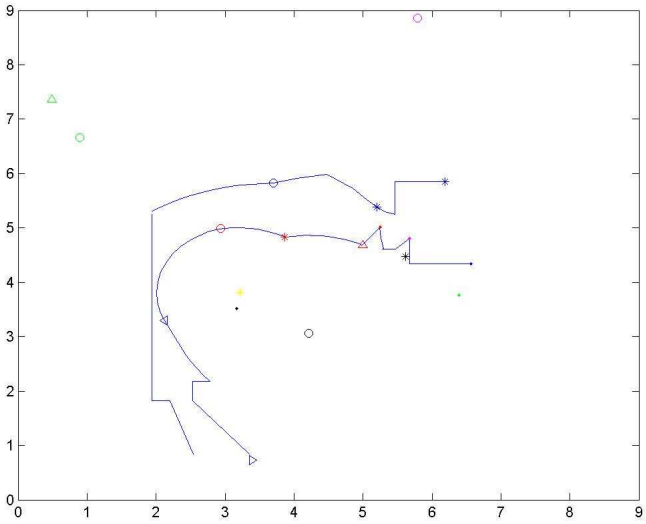
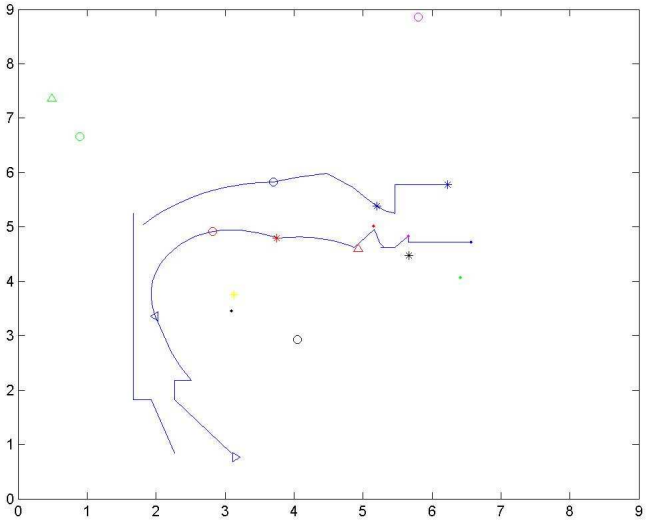


Figure 5.8: Graphical representation of Childers' vocal tract model with landmarks marking out the critical anatomy structures and articulatory positions that determine the configuration of the vocal tract. The critical anatomical structures include M (posterior end of the hard palate), N (anterior end of the hard palate) and U (upper incisor). The critical articulatory positions include H (hyoid bone), DL (separation landmark of pharynx and oral cavity), B (tongue body), T (tongue tip), JAW (lower incisor), L5 (upper lip), L7 (lower lip), V (highest position of the velum when it contacts the posterior wall of the pharynx), V' (velum) and *tongc* (center of tongue body).

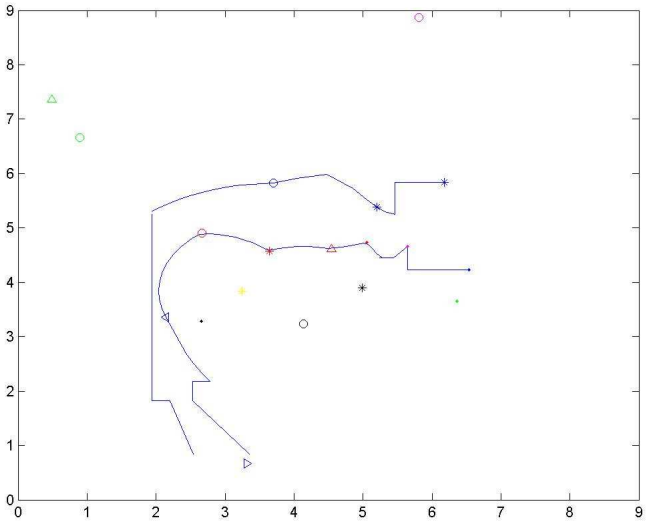


(a) /a/ from /aba/

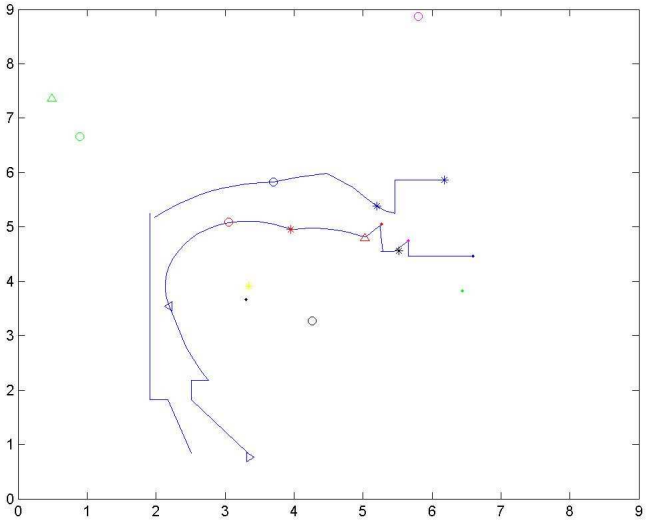


(b) /a/ from /ama/

Figure 5.9: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

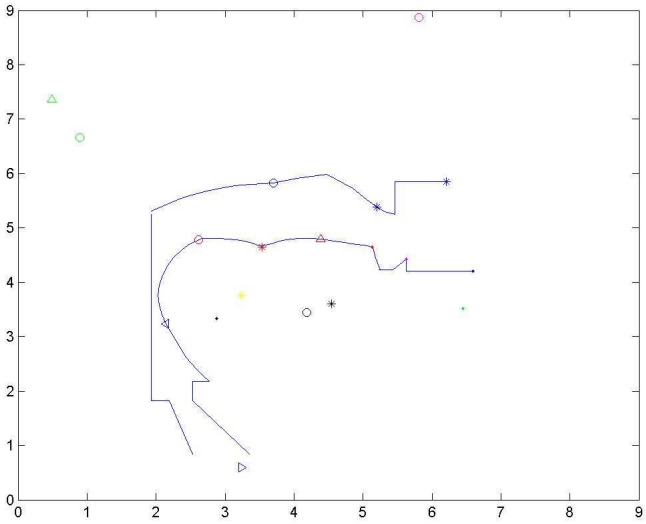


(a) /a/ from /ada/

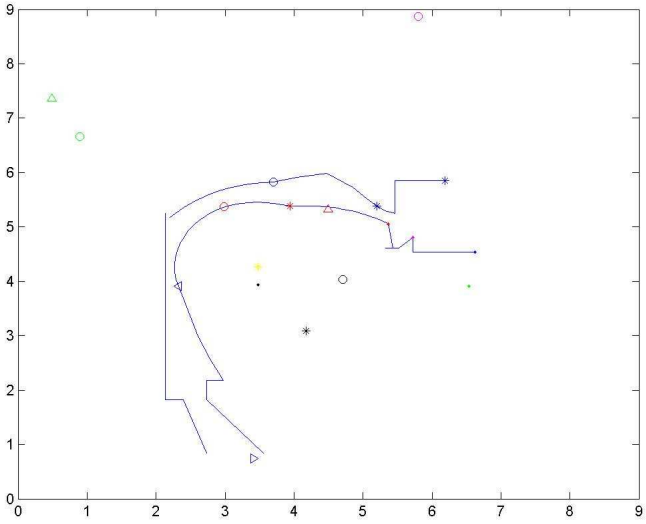


(b) /a/ from /ana/

Figure 5.10: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

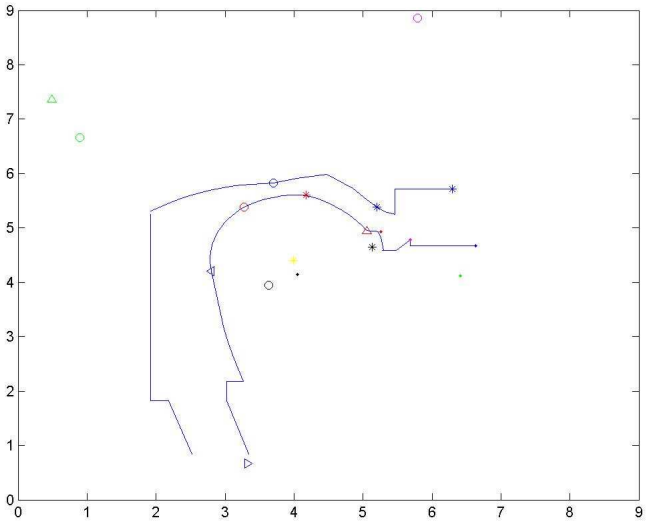


(a) /a/ from /aga/

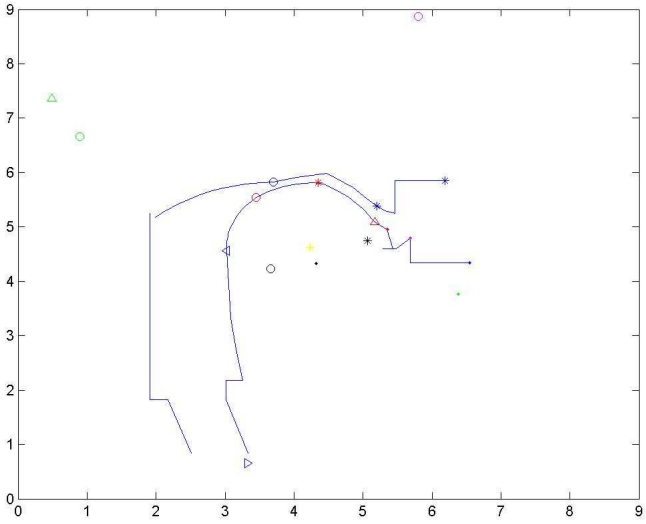


(b) /a/ from /aŋa/

Figure 5.11: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

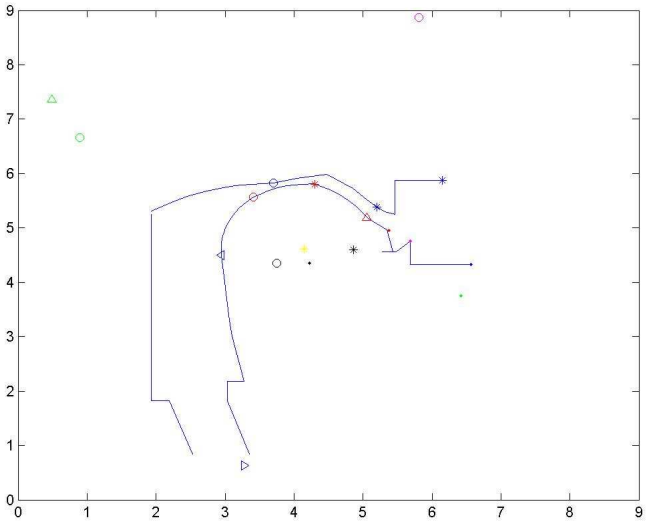


(a) /i/ from /bip/

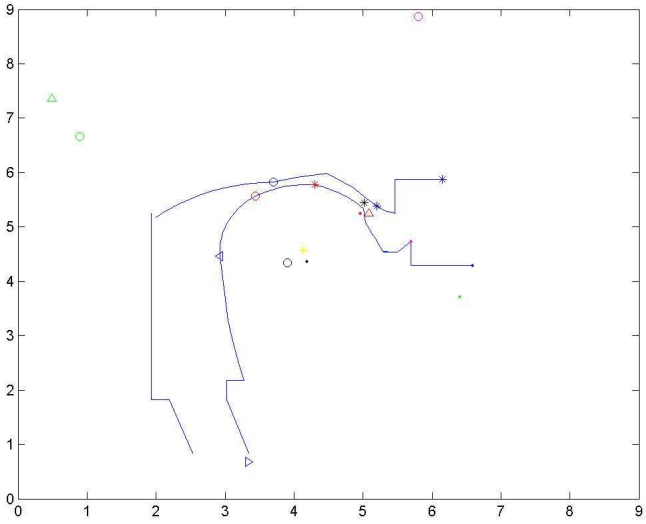


(b) /i/ from /bim/

Figure 5.12: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

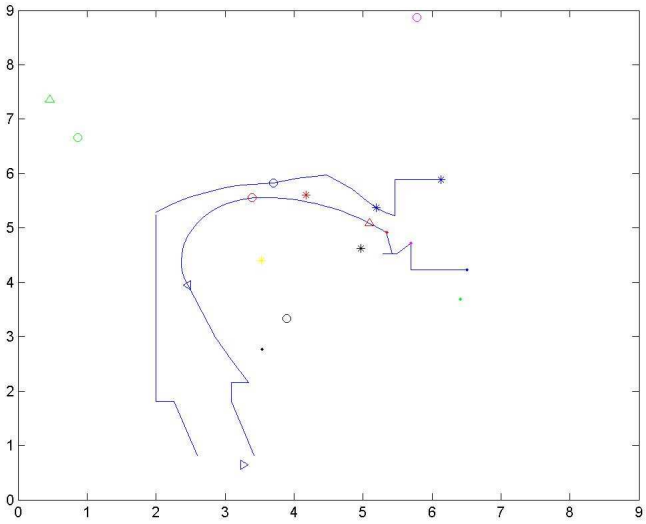


(a) /i/ from /dit/

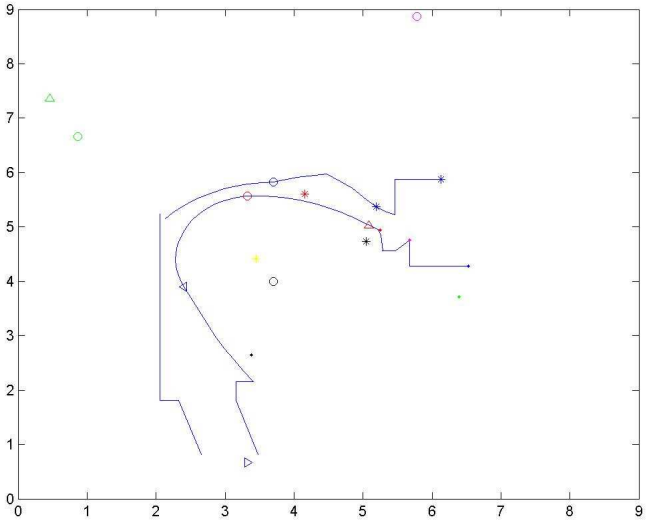


(b) /i/ from /din/

Figure 5.13: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

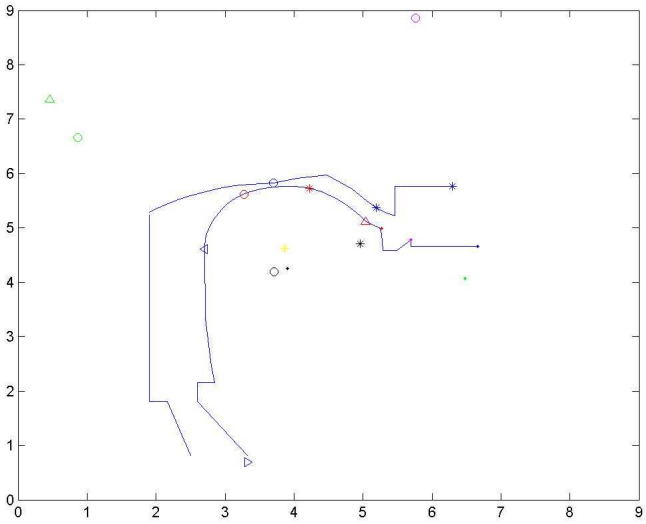


(a) /i/ from /kik/

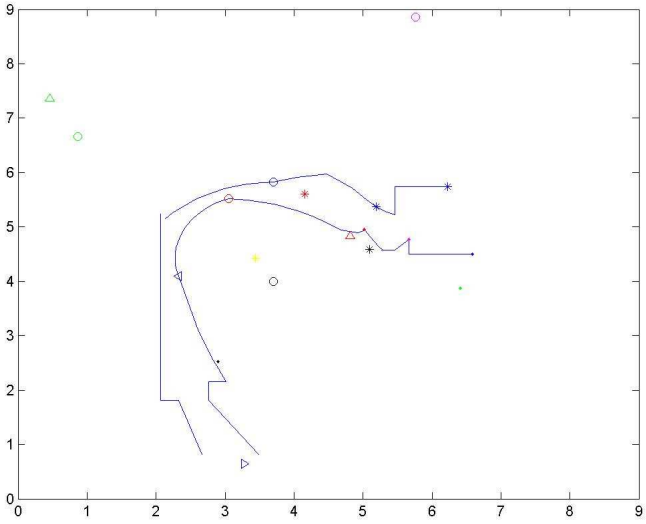


(b) /i/ from /kiŋ/

Figure 5.14: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

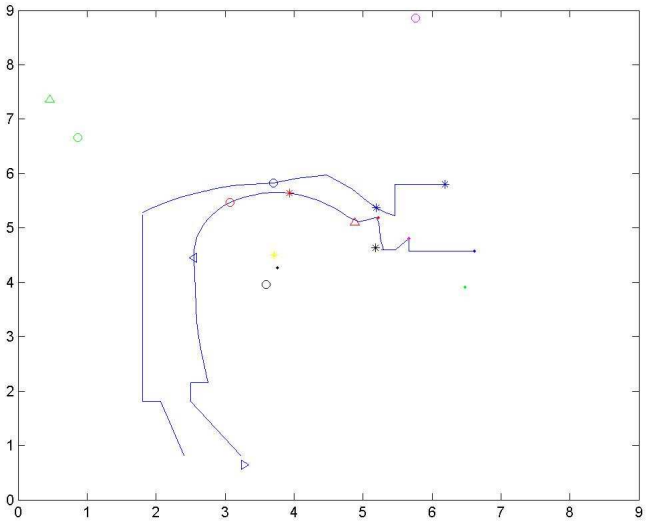


(a) /u/ from /pub/

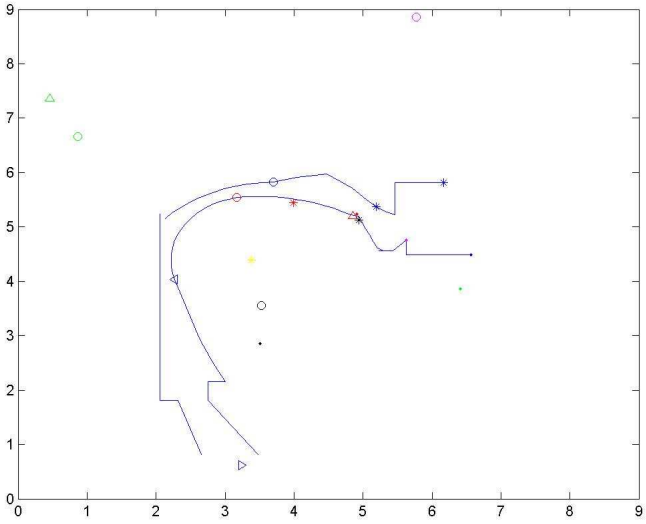


(b) /u/ from /pum/

Figure 5.15: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

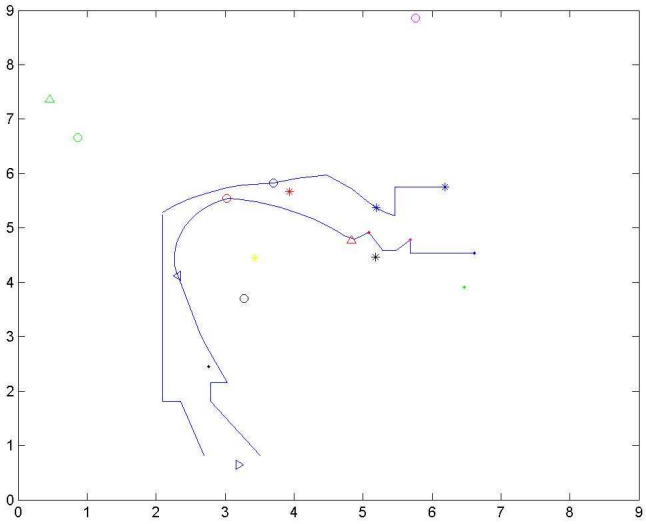


(a) /u/ from /tud/

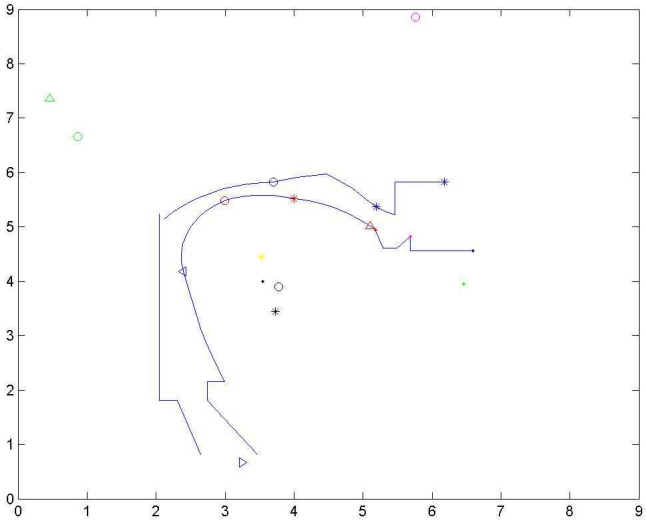


(b) /u/ from /tun/

Figure 5.16: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

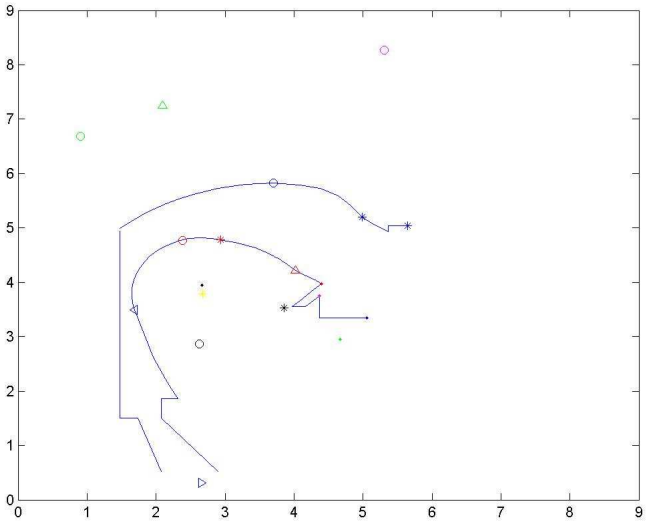


(a) /u/ from /kuk/

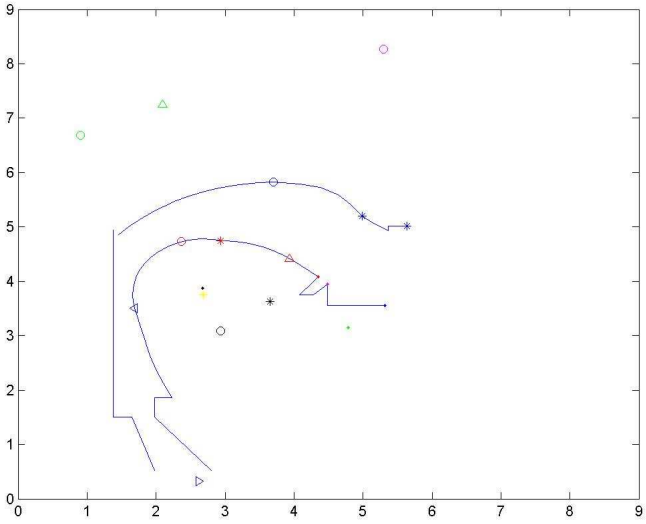


(b) /u/ from /kuŋ/

Figure 5.17: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 1. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

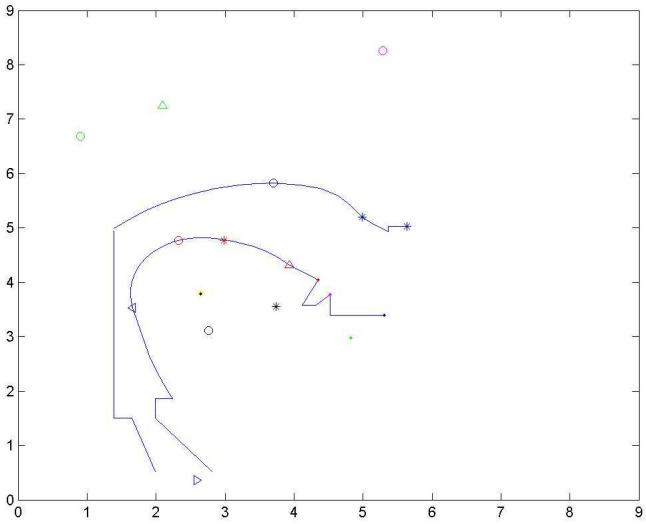


(a) /a/ from /pab/

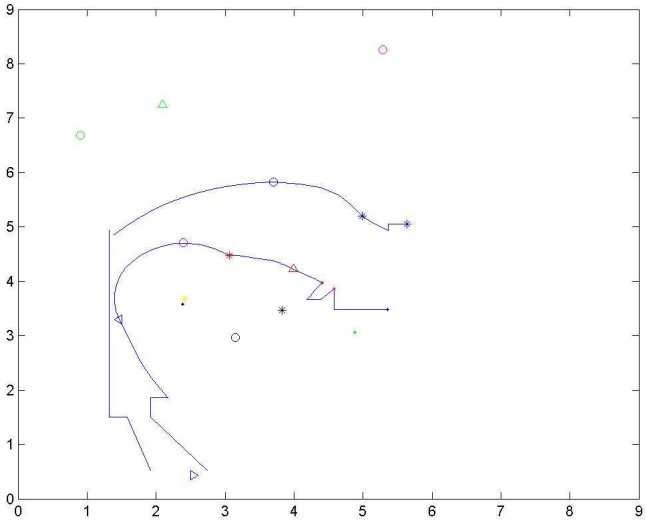


(b) /a/ from /pam/

Figure 5.18: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

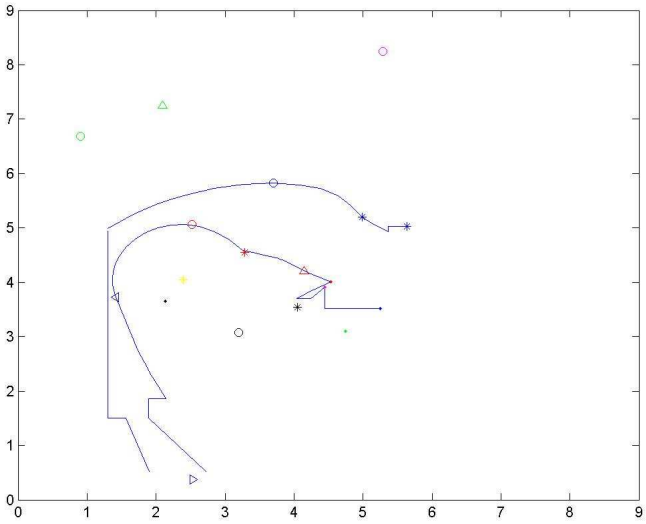


(a) /a/ from /tad/

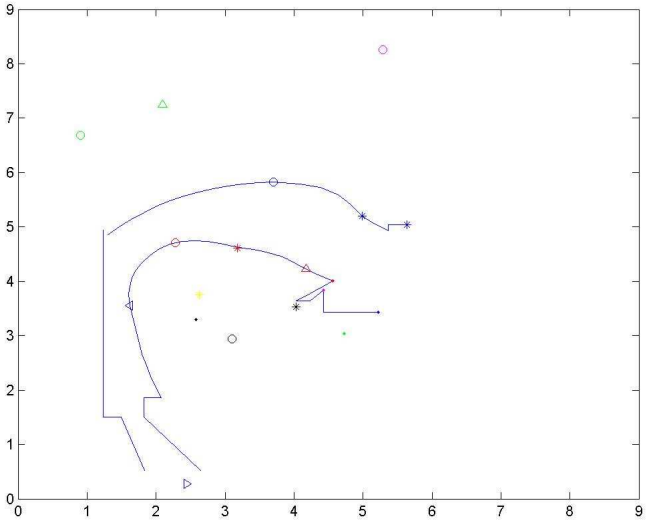


(b) /a/ from /tan/

Figure 5.19: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

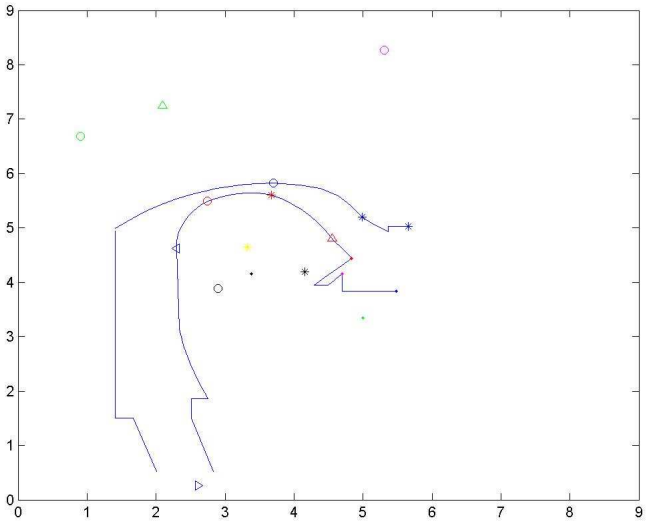


(a) /a/ from /kak/

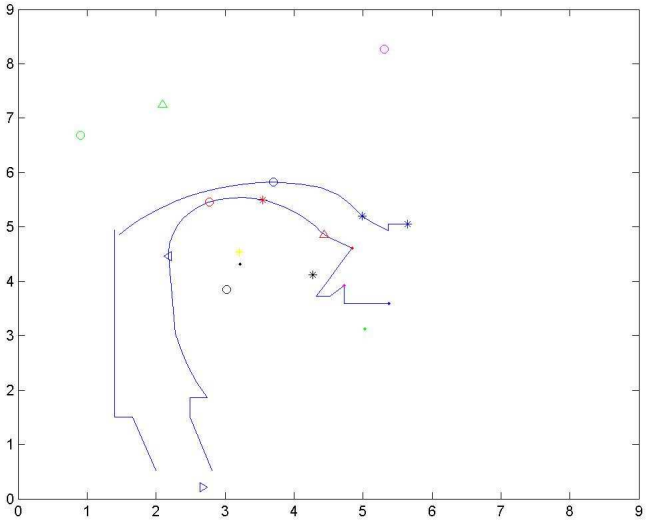


(b) /a/ from /kaŋ/

Figure 5.20: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

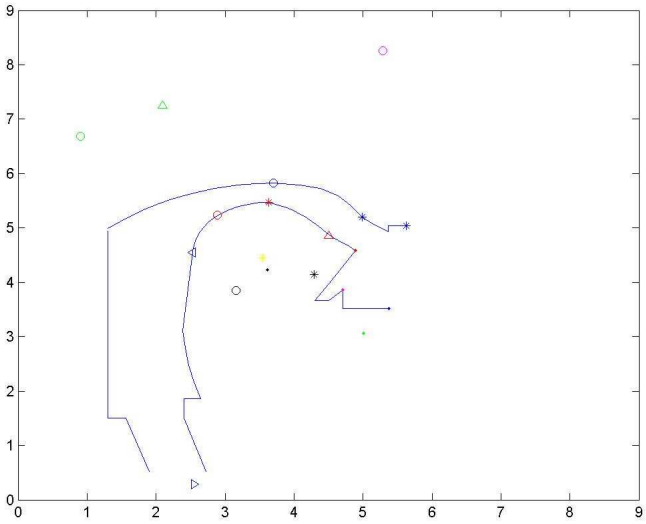


(a) /i/ from /ibi/

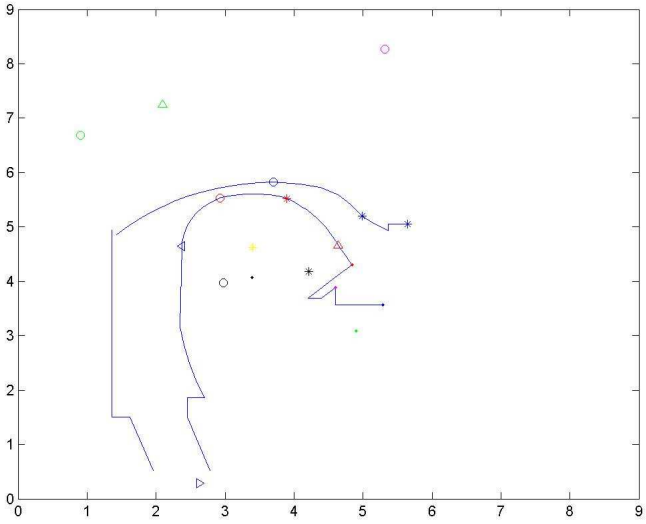


(b) /i/ from /imi/

Figure 5.21: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

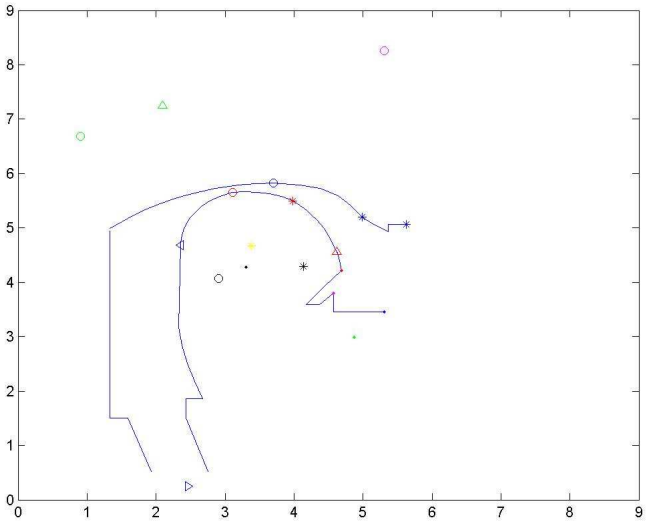


(a) /i/ from /idi/

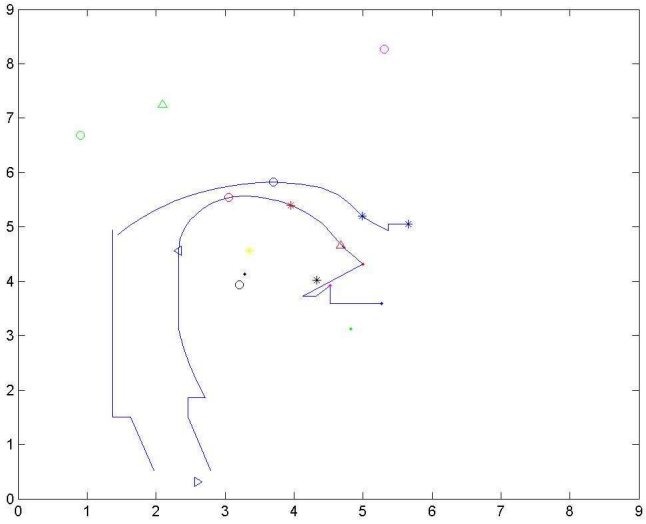


(b) /i/ from /ini/

Figure 5.22: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

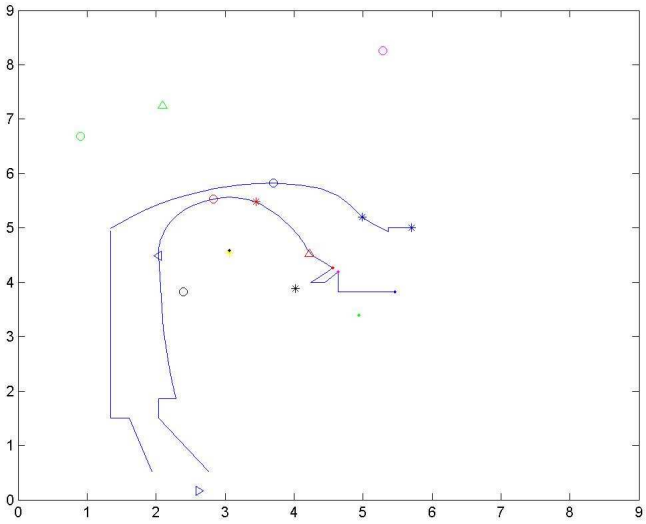


(a) /i/ from /igi/

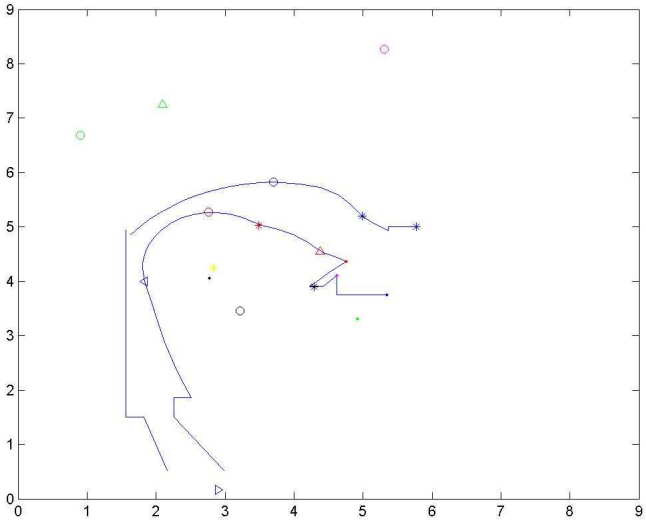


(b) /i/ from /iŋi/

Figure 5.23: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

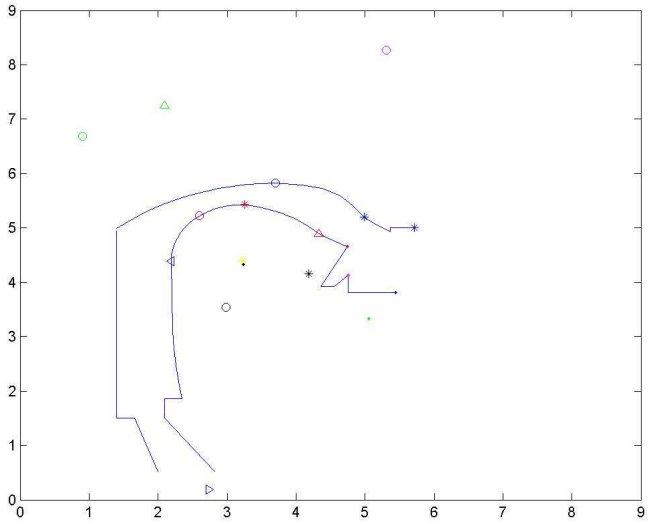


(a) /u/ from /bub/

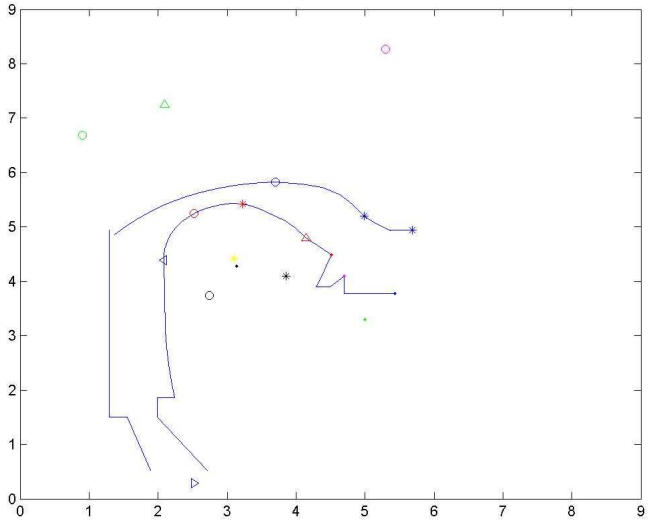


(b) /a/ from /bum/

Figure 5.24: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

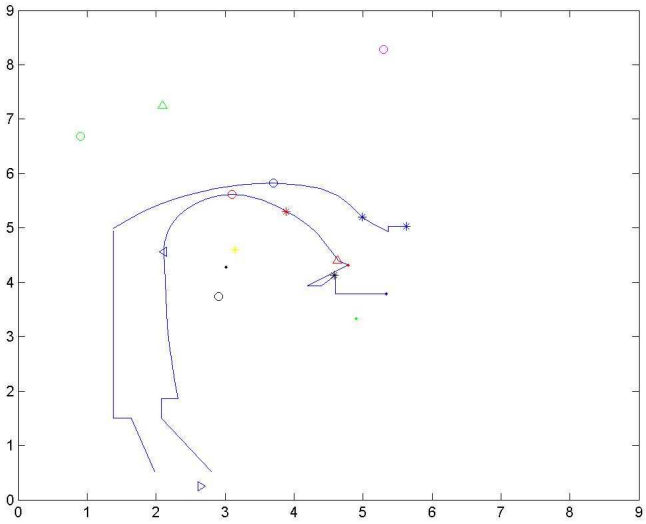


(a) /u/ from /dud/

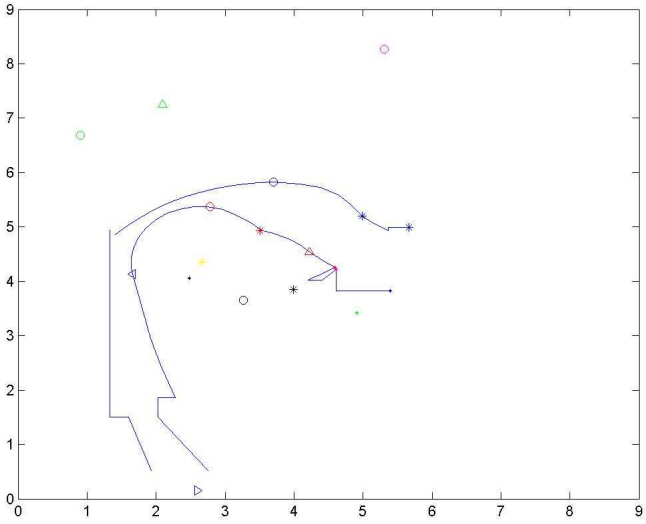


(b) /u/ from /dun/

Figure 5.25: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).



(a) /u/ from /gug/



(b) /u/ from /guŋ/

Figure 5.26: Midsagittal configuration of the vocal tract estimated from the EMA measurements of Speaker 2. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), TB (upper triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL, respectively. The asterisk in the hard palate region represents UI and the dot posterior to LL corresponds to LI. The upper triangle and the two circles above the palate correspond to the three reference structures (LT, RT, Nose). In addition to the EMA measurements, the critical articulatory landmarks used in construction of the model are also marked in the figure, including the center of each part of the tongue (dot: *tongc*; asterisks: *c1b*, *c2b*; circle: *c1t*) and the position of DL (lower triangle).

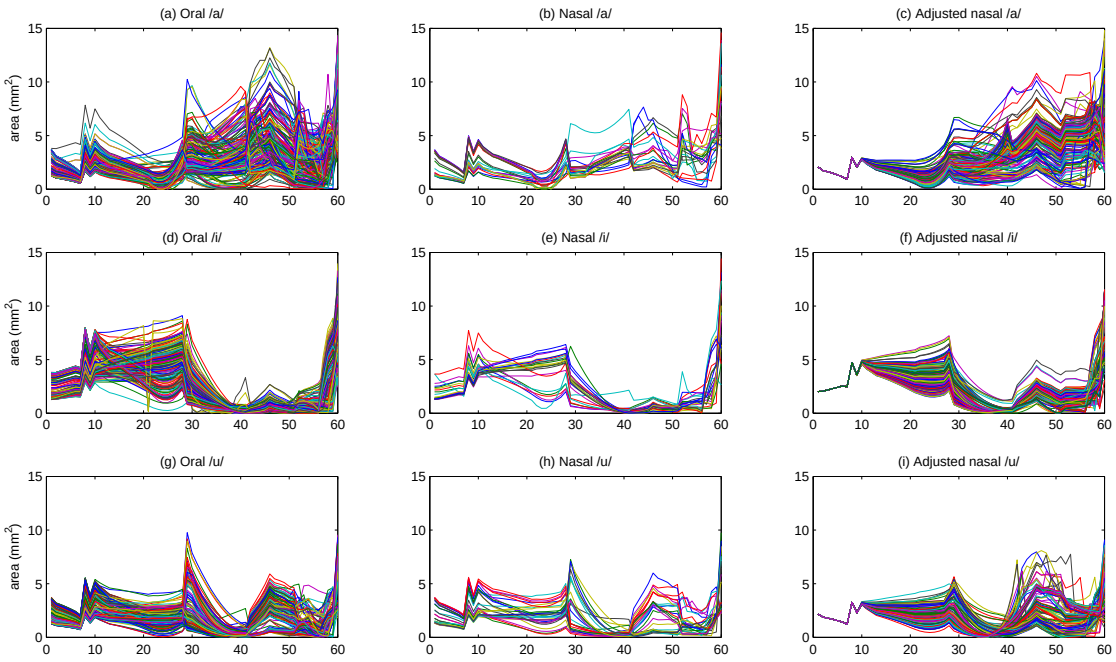


Figure 5.27: The area functions computed based on the estimated midsagittal configurations of the vocal tract and the speaker-adaptive PONM for model of Speaker 1.

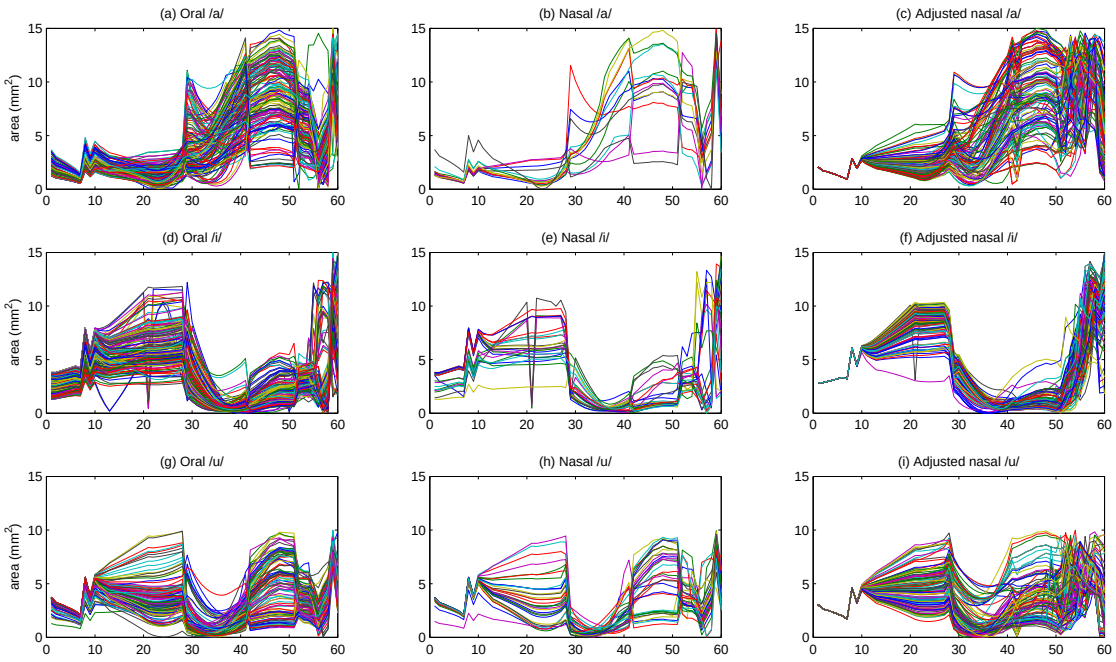


Figure 5.28: The area functions computed based on the estimated midsagittal configurations of the vocal tract and the speaker-adaptive PONM for model of Speaker 2.

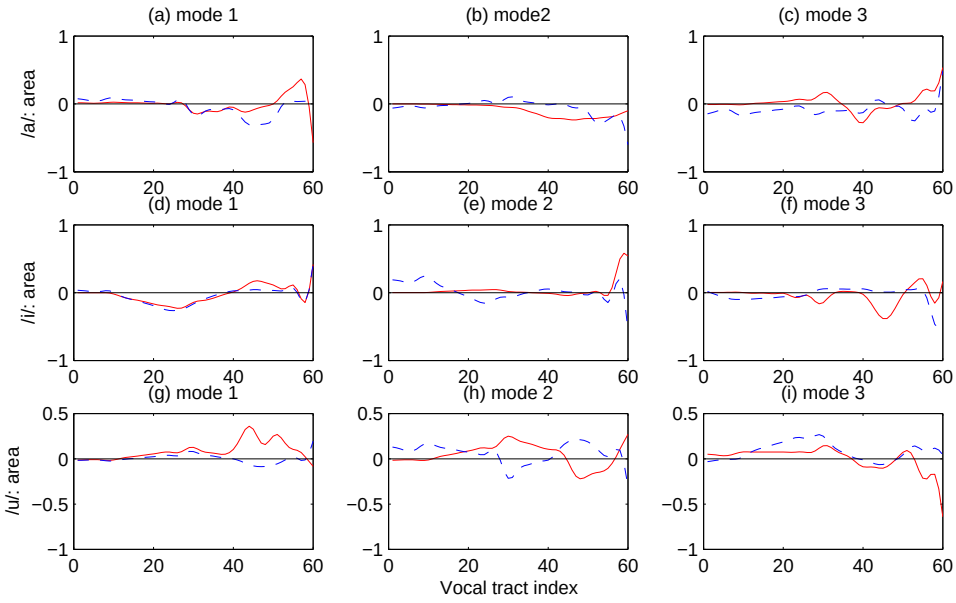


Figure 5.29: The first three principal orthogonal articulatory modes of Speaker 1 for /a/ in (a)–(c), /i/ in (d)–(f), and /u/ in (g)–/i/. The red solid lines correspond to the articulatory modes for NA and the blue dashed lines correspond to the articulatory modes for N. The x-axis is the index from the glottis and the y-axis is area difference between NA and O or between N and O.

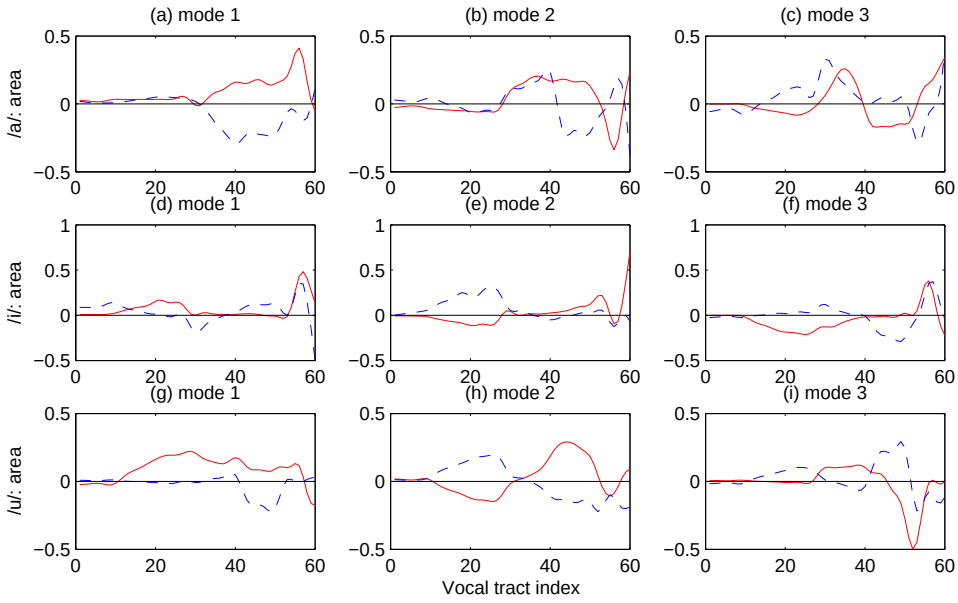


Figure 5.30: The first three principal orthogonal articulatory modes of Speaker 2 for /a/ in (a)–(c), /i/ in (d)–(f), and /u/ in (g)–/i/. The red solid lines correspond to the articulatory modes for NA and the blue dashed lines correspond to the articulatory modes for N. The x-axis is the index from the glottis and the y-axis is area difference between NA and O or between N and O.

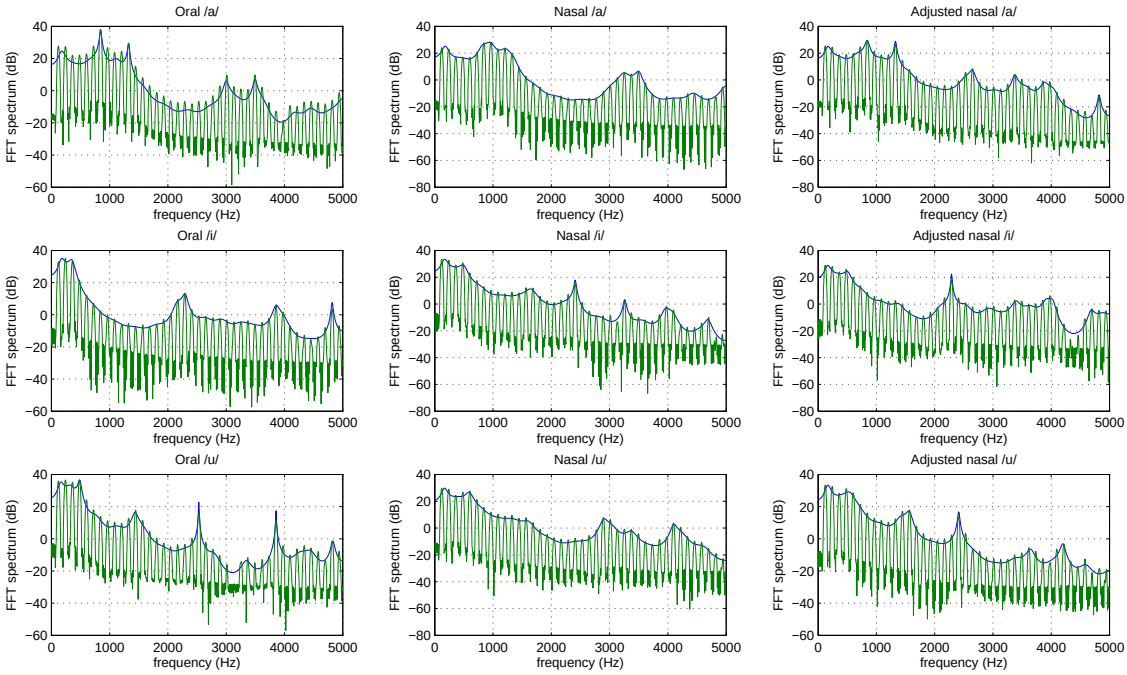


Figure 5.31: Spectra of the following synthetic vowel samples: oral /a/, nasal /a/ and adjusted nasal /a/ from left to right in the first row; oral /i/, nasal /i/ and adjusted nasal /i/ from left to right in the second row; oral /u/, nasal /u/ and adjusted nasal /u/ from left to right in the third row.

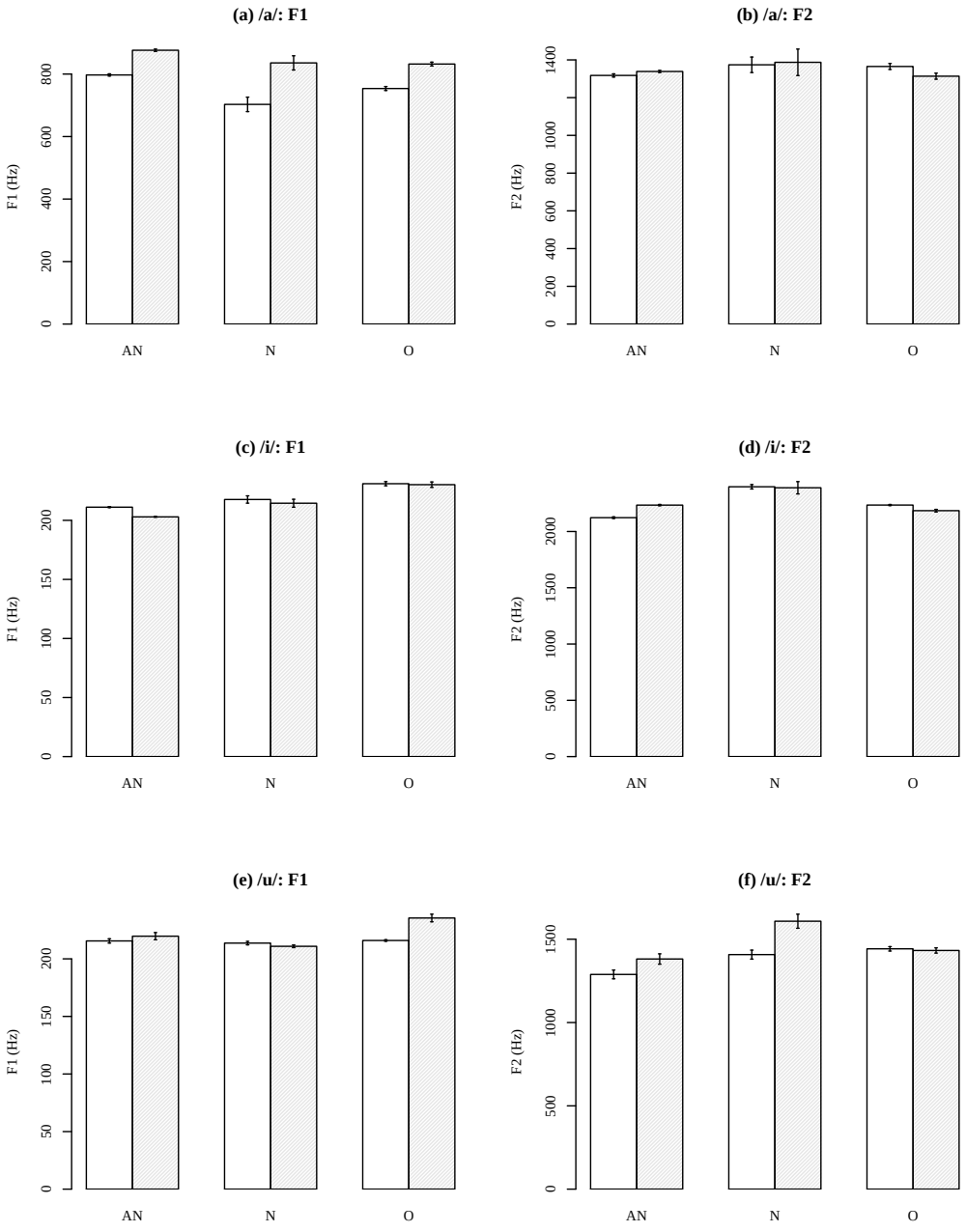


Figure 5.32: Barplots of the F1 and F2 of the synthetic vowels. (a) F1 of /a/, (b) F2 of /a/, (c) F1 of /i/, (d) F2 of /i/, (e) F1 of /u/, (f) F2 of /u/. The white and shaded bars correspond to the vowel samples synthesized with the models of Speaker 1 and 2, respectively. The three groups "NA", "N" and "O" correspond to nasal vowels with articulatory adjustment, nasal vowels and oral vowels, respectively.

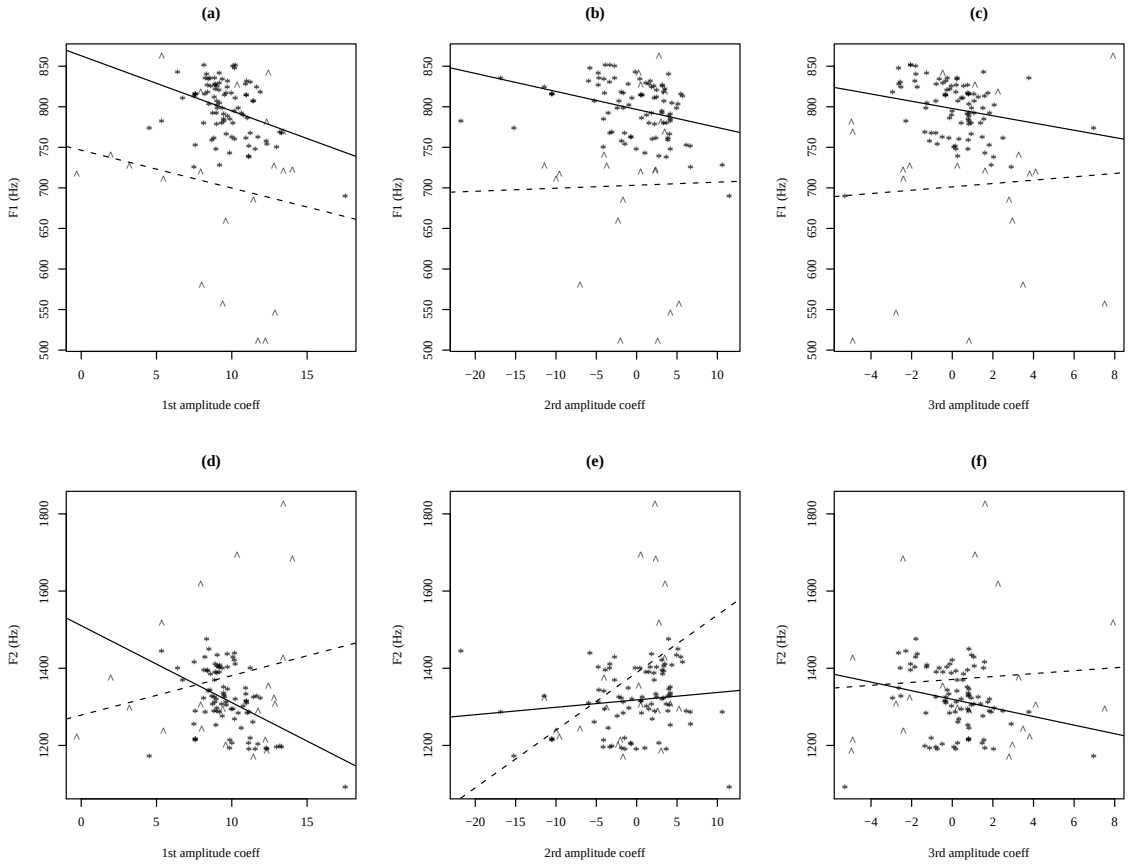


Figure 5.33: Scatterplots of formant frequencies versus amplitude coefficients of the orthogonal modes for /a/ synthesized with the model of Speaker 1. (a) F1–pp1, (b) F1–pp2, (c) F1–pp3, (d) F2–pp1, (e) F2–pp2, (f) F2–pp3. The asterisks and upper triangles correspond to the data of NA and N, respectively. The solid and dashed lines are the linear regression lines fitted to the data of NA and N, respectively.

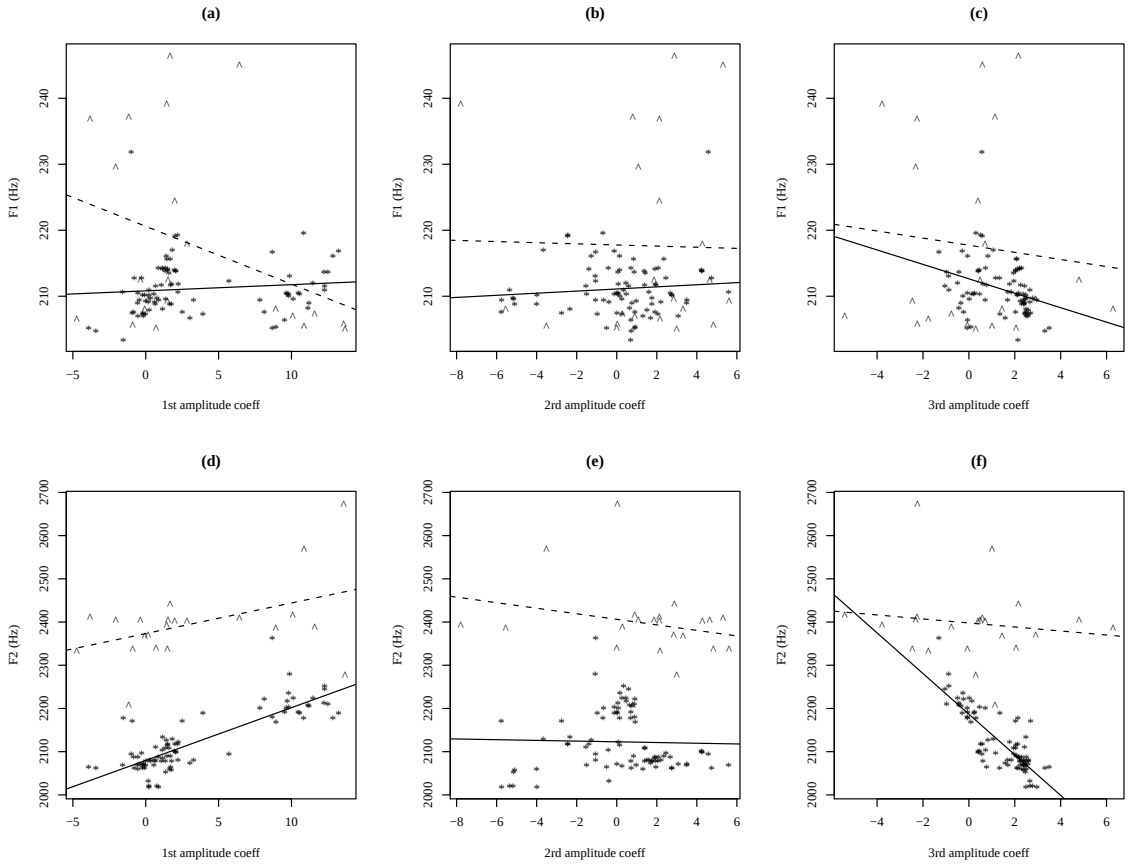


Figure 5.34: Scatterplots of formant frequencies versus amplitude coefficients of the orthogonal modes for /i/ synthesized with the model of Speaker 1. (a) F1–pp1, (b) F1–pp2, (c) F1–pp3, (d) F2–pp1, (e) F2–pp2, (f) F2–pp3. The asterisks and upper triangles correspond to the data of NA and N, respectively. The solid and dashed lines are the linear regression lines fitted to the data of NA and N, respectively.

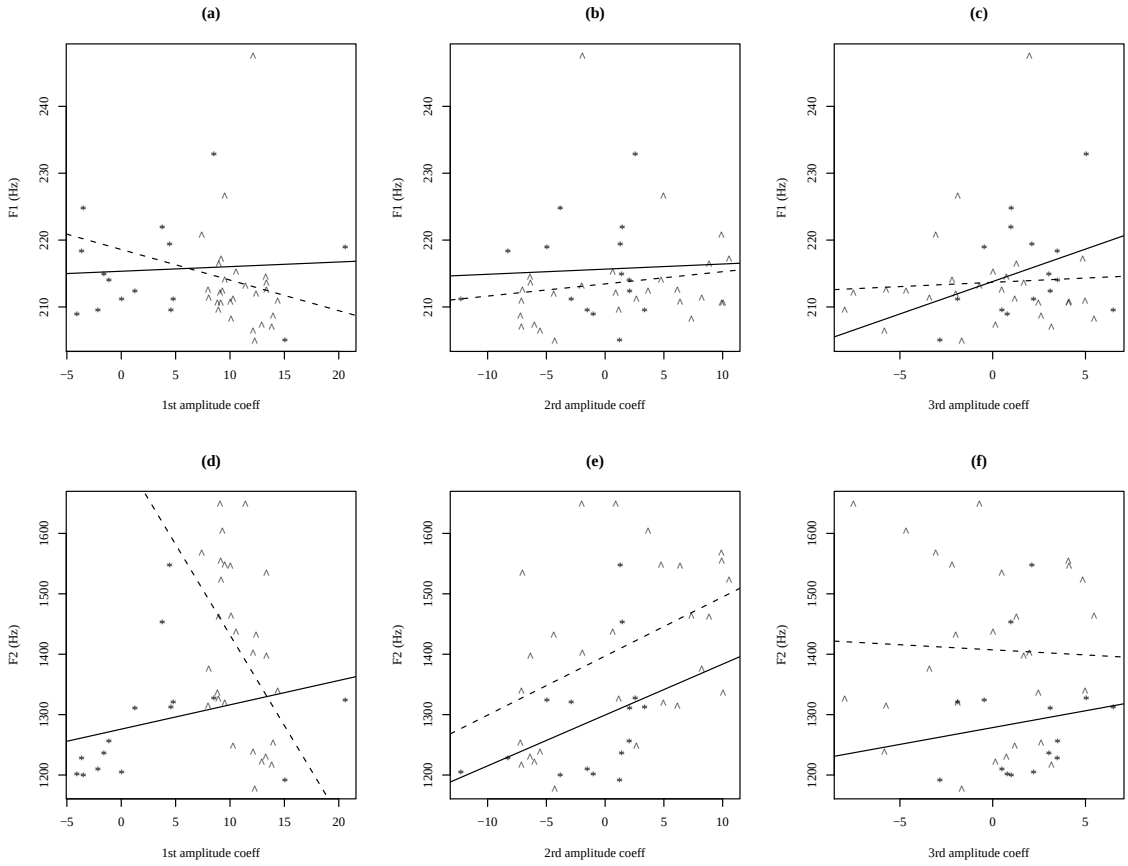


Figure 5.35: Scatterplots of formant frequencies versus amplitude coefficients of the orthogonal modes for /u/ synthesized with the model of Speaker 1. (a) F1–pp1, (b) F1–pp2, (c) F1–pp3, (d) F2–pp1, (e) F2–pp2, (f) F2–pp3. The asterisks and upper triangles correspond to the data of NA and N, respectively. The solid and dashed lines are the linear regression lines fitted to the data of NA and N, respectively.

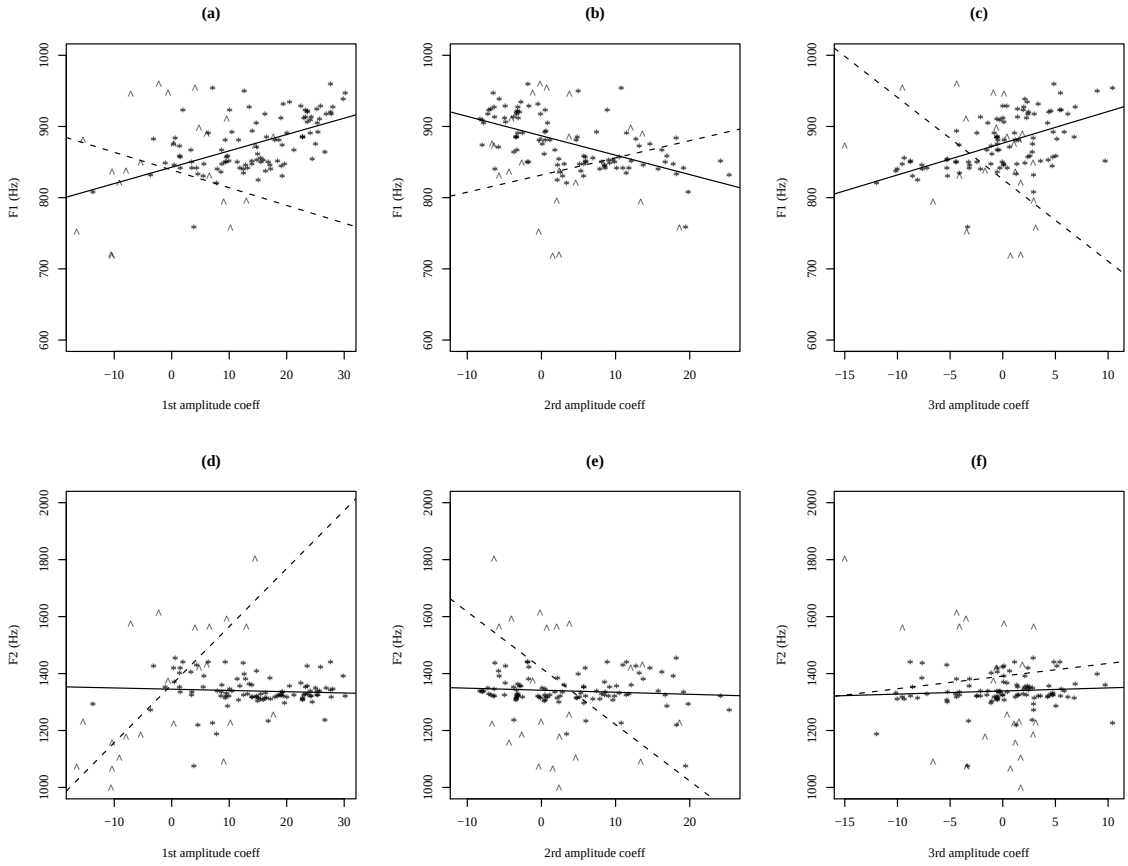


Figure 5.36: Scatterplots of formant frequencies versus amplitude coefficients of the orthogonal modes for /a/ synthesized with the model of Speaker 2. (a) F1–pp1, (b) F1–pp2, (c) F1–pp3, (d) F2–pp1, (e) F2–pp2, (f) F2–pp3. The asterisks and upper triangles correspond to the data of NA and N, respectively. The solid and dashed lines are the linear regression lines fitted to the data of NA and N, respectively.

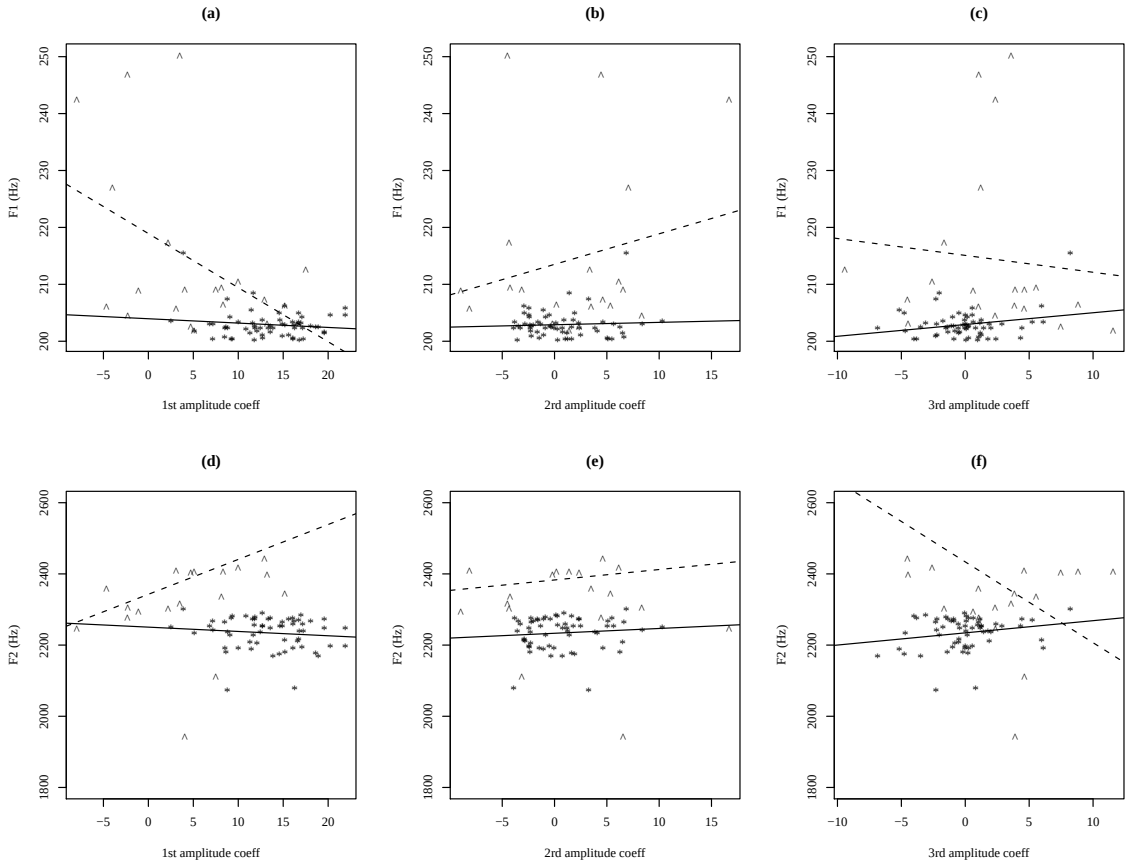


Figure 5.37: Scatterplots of formant frequencies versus amplitude coefficients of the orthogonal modes for /i/ synthesized with the model of Speaker 2. (a) F1–pp1, (b) F1–pp2, (c) F1–pp3, (d) F2–pp1, (e) F2–pp2, (f) F2–pp3. The asterisks and upper triangles correspond to the data of NA and N, respectively. The solid and dashed lines are the linear regression lines fitted to the data of NA and N, respectively.

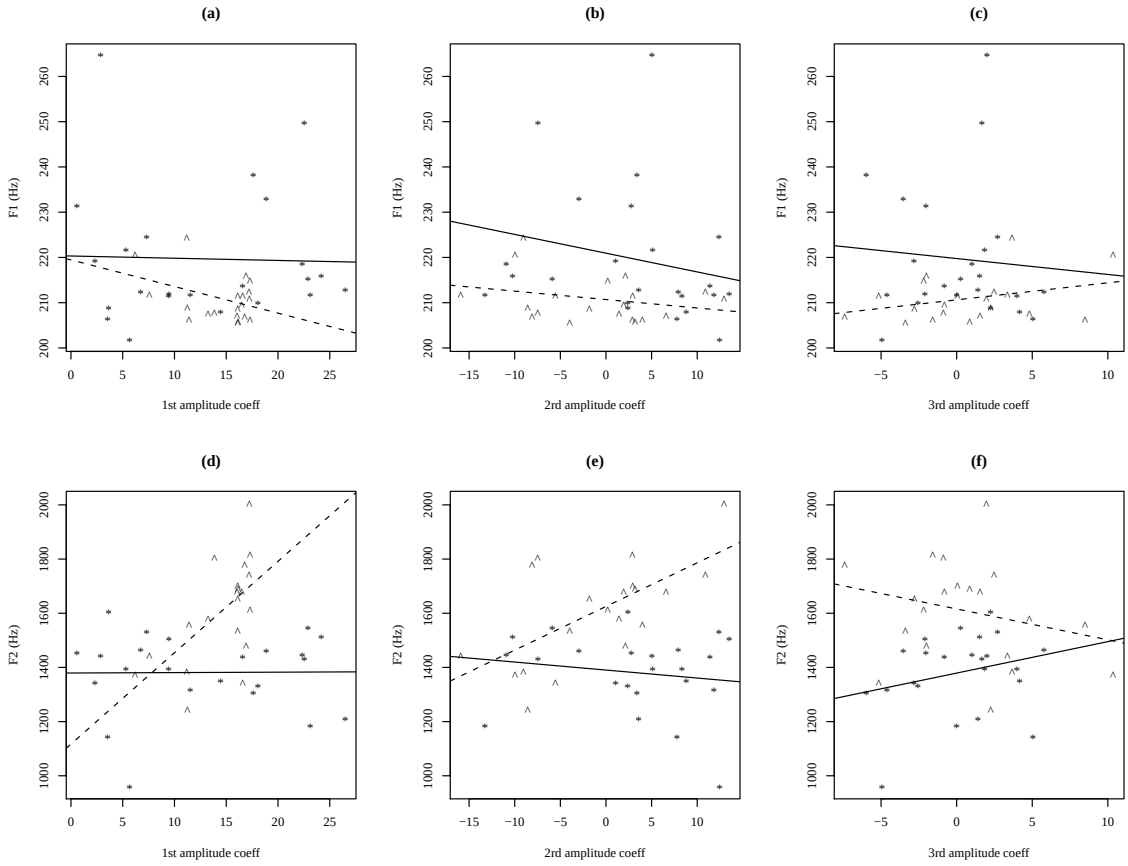


Figure 5.38: Scatterplots of formant frequencies versus amplitude coefficients of the orthogonal modes for /u/ synthesized with the model of Speaker 2. (a) F1–pp1, (b) F1–pp2, (c) F1–pp3, (d) F2–pp1, (e) F2–pp2, (f) F2–pp3. The asterisks and upper triangles correspond to the data of NA and N, respectively. The solid and dashed lines are the linear regression lines fitted to the data of NA and N, respectively.

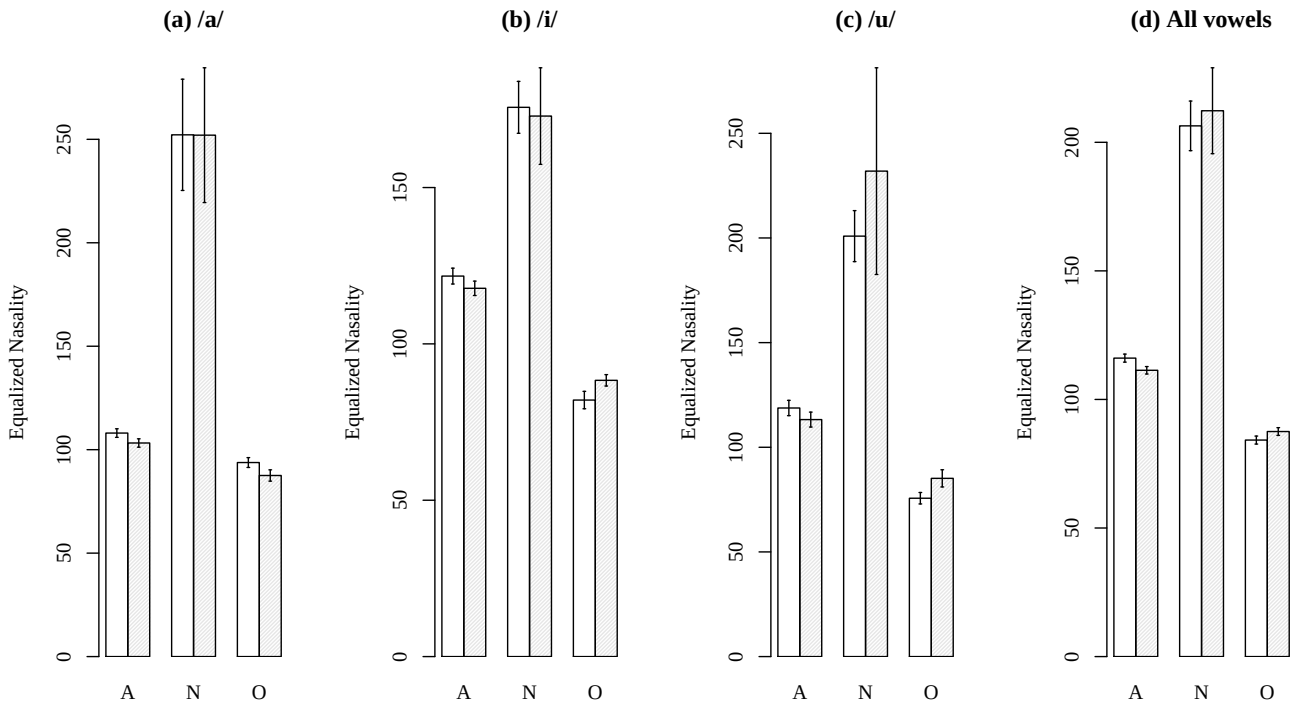


Figure 5.39: Barplots of the equalized nasality scores with respect to vowel type (O: oral vowels, N: nasal vowels, A: nasal vowels with articulatory adjustment). (a)–(c) show the nasality scores for individual vowels /a/, /i/ and /u/, respectively; (d) shows the average of nasality scores across the three vowels /a/, /i/ and /u/. The white and shaded bars correspond to Speaker 1 and 2, respectively.

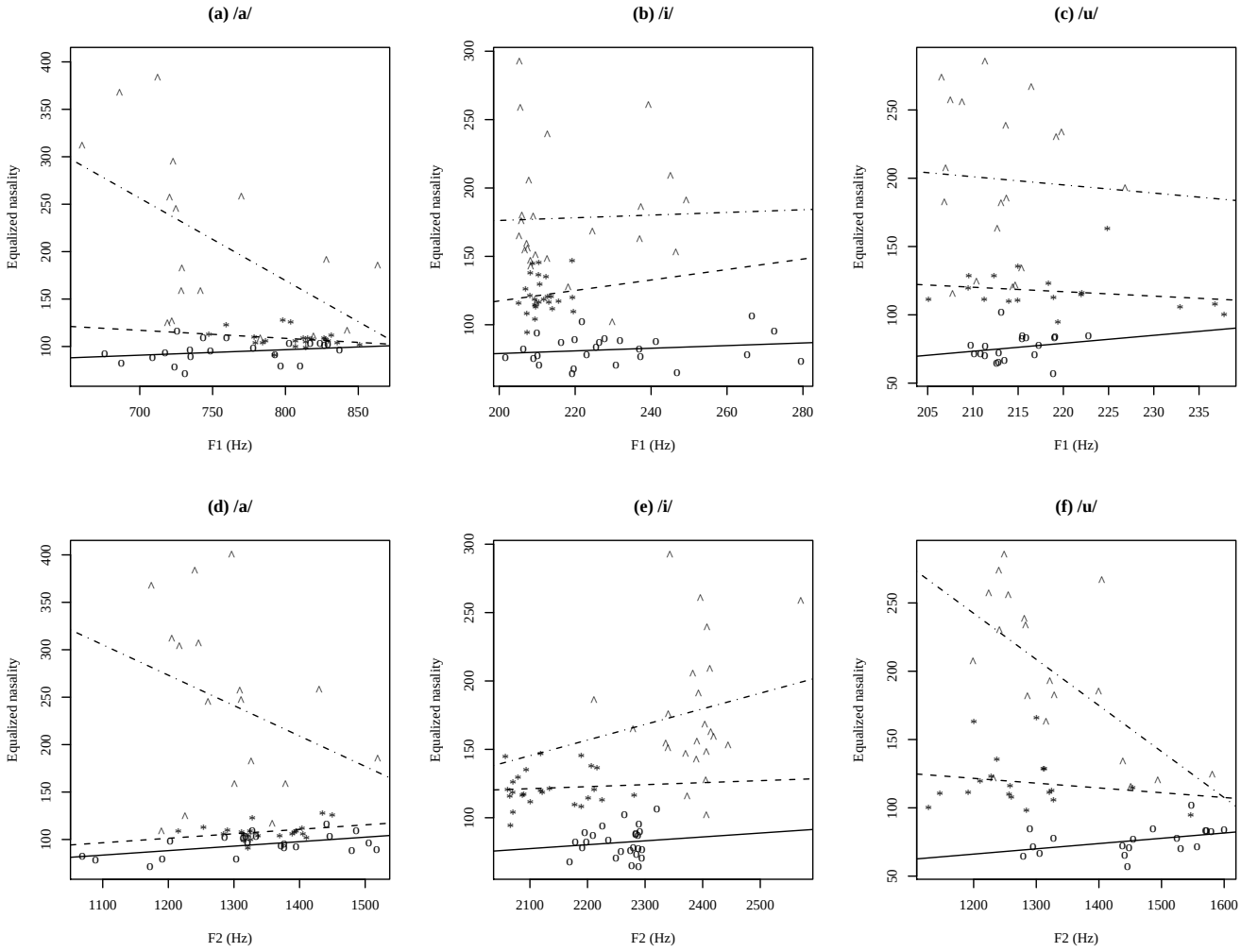


Figure 5.40: Scatterplots of the equalized nasality scores versus formant frequencies for Speaker 1. (a) nasality–F1 for /a/, (b) nasality–F1 for /i/, (c) nasality–F1 for /u/, (d) nasality–F2 for /a/, (e) nasality–F2 for /i/, (f) nasality–F2 for /u/. The circles, upper triangles and asterisks correspond to O, N and NA, respectively. The solid, dot-dashed and dashed lines are the linear regression fits to the data of O, N and NA, respectively.

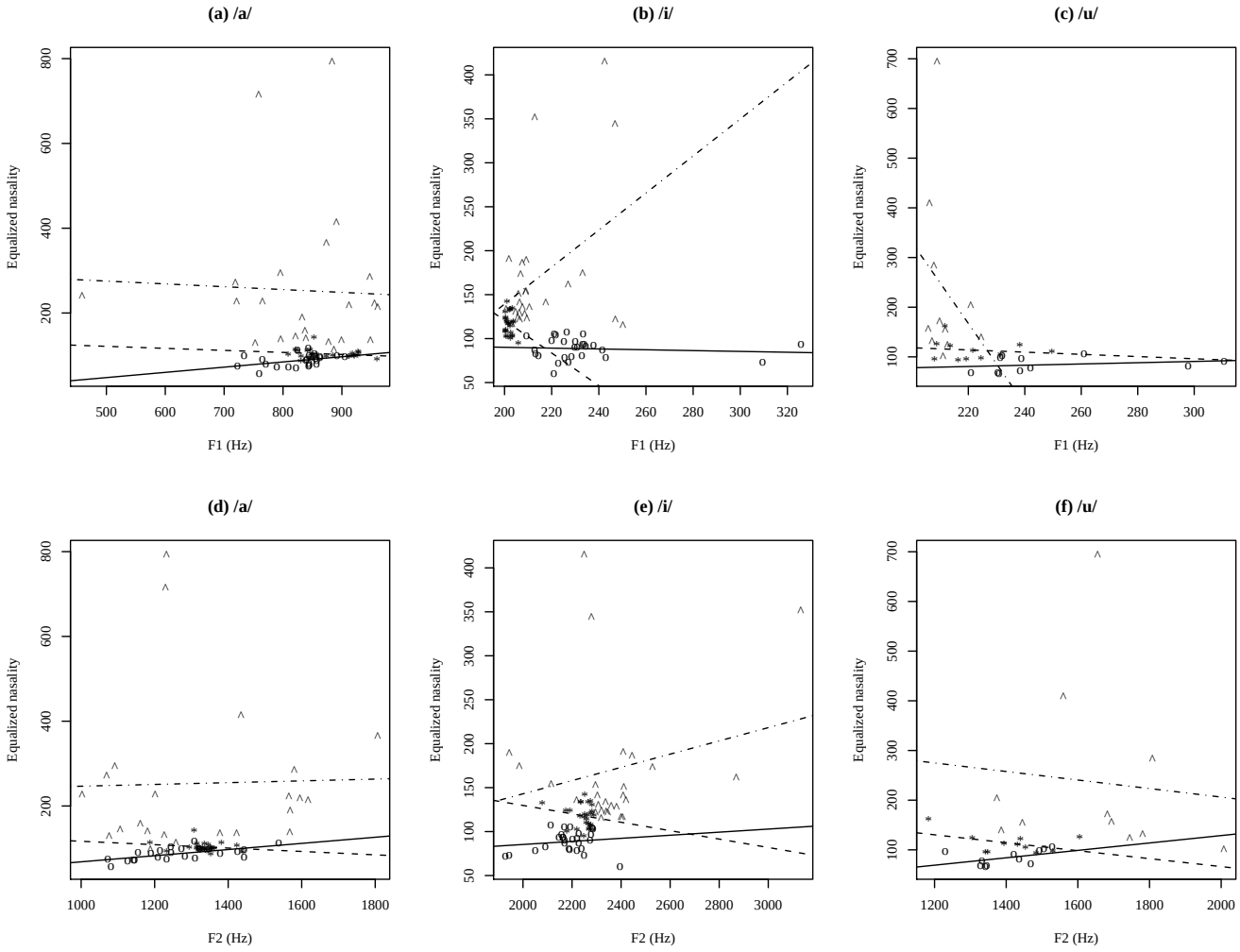


Figure 5.41: Scatterplots of the equalized nasality scores versus formant frequencies for Speaker 2. (a) nasality–F1 for /a/, (b) nasality–F1 for /i/, (c) nasality–F1 for /u/, (d) nasality–F2 for /a/, (e) nasality–F2 for /i/, (f) nasality–F2 for /u/. The circles, upper triangles and asterisks correspond to O, N and NA, respectively. The solid, dot-dashed and dashed lines are the linear regression fits to the data of O, N and NA, respectively.

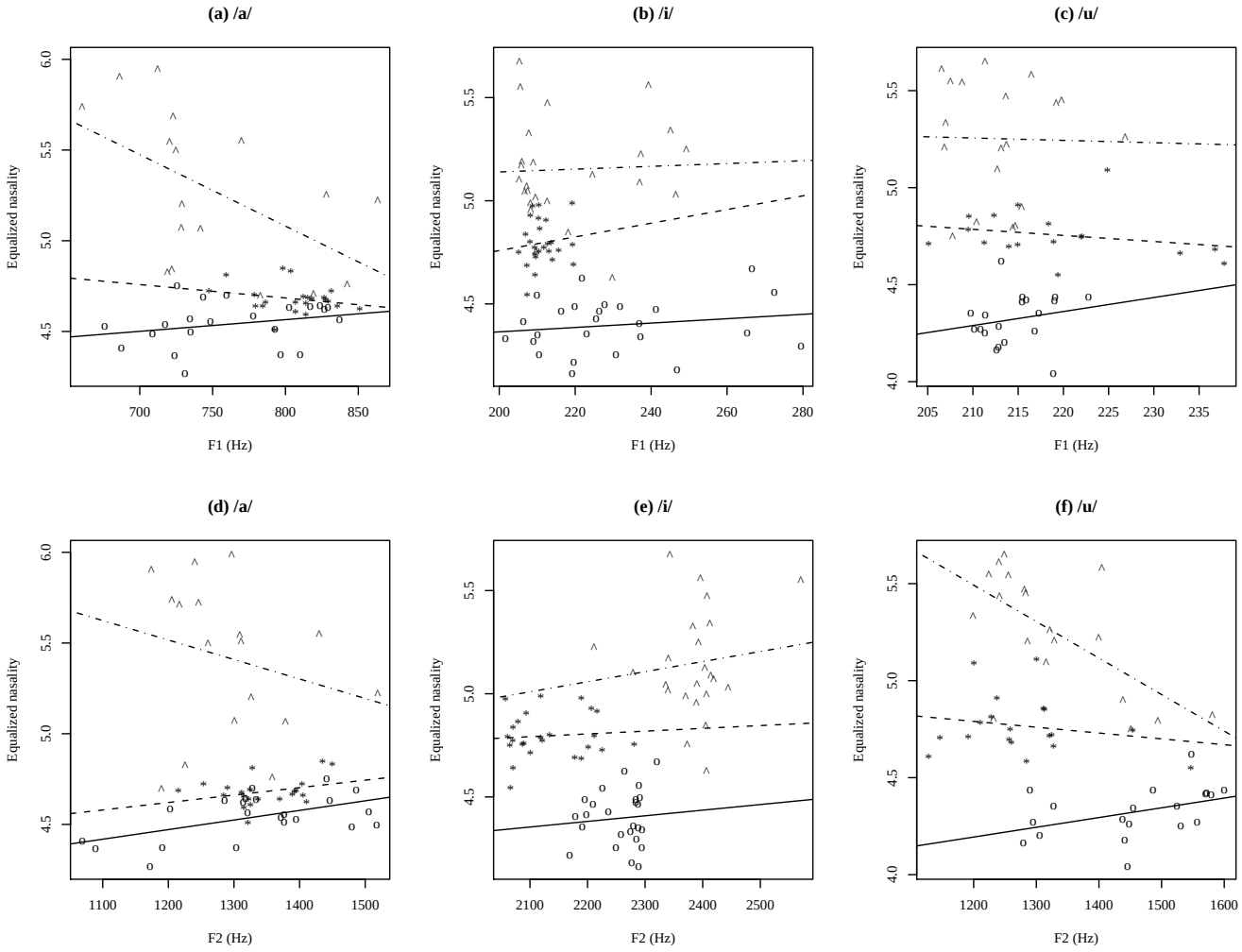


Figure 5.42: Scatterplots of the logarithm of nasality scores versus formant frequencies for Speaker 1. (a) log(nasality)–F1 for /a/, (b) log(nasality)–F1 for /i/, (c) log(nasality)–F1 for /u/, (d) log(nasality)–F2 for /a/, (e) log(nasality)–F2 for /i/, (f) log(nasality)–F2 for /u/. The circles, upper triangles and asterisks correspond to O, N and NA, respectively. The solid, dot-dashed and dashed lines are the linear regression fits to the data of O, N and NA, respectively.

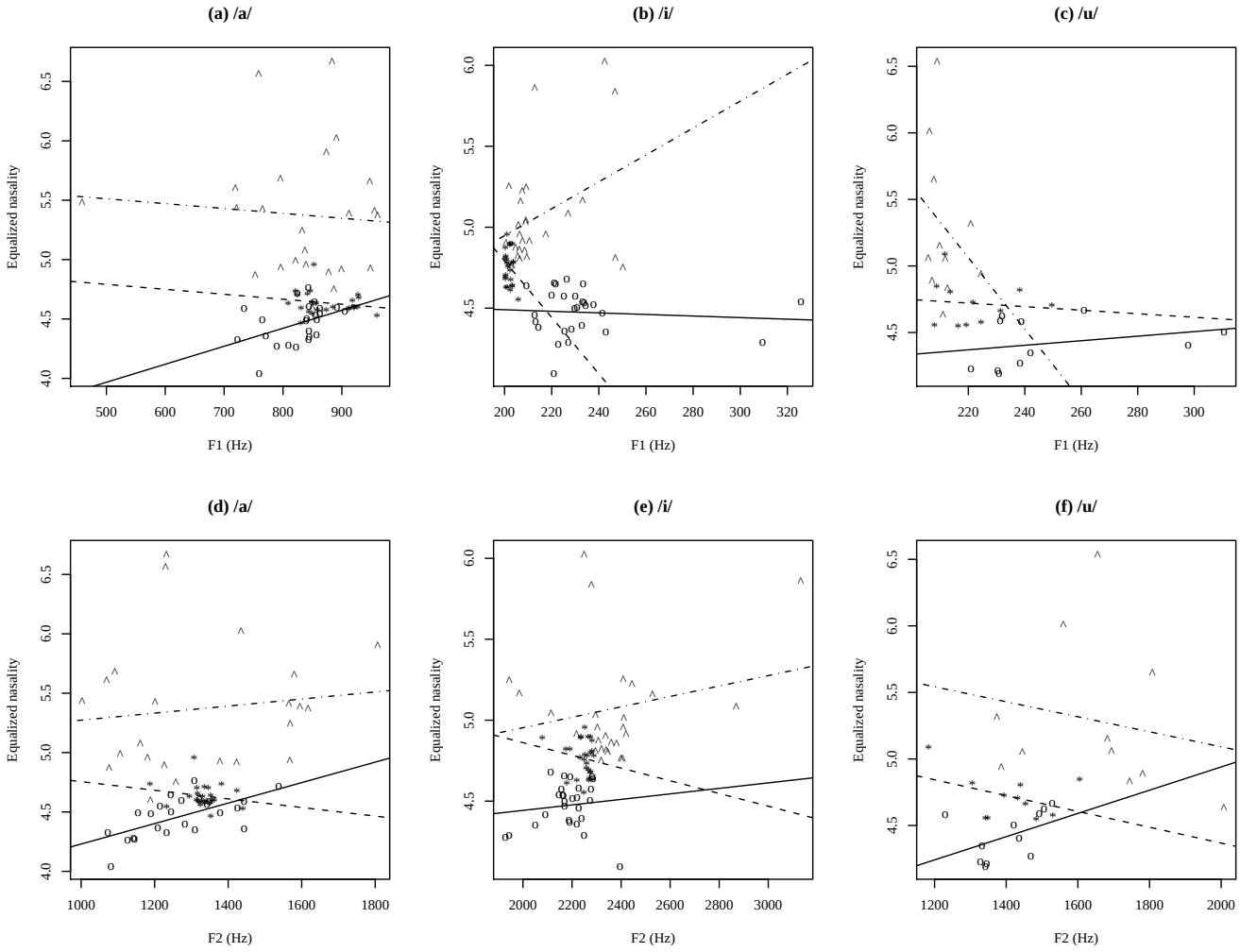


Figure 5.43: Scatterplots of the logarithm of nasality scores versus formant frequencies for Speaker 2. (a) log(nasality)–F1 for /a/, (b) log(nasality)–F1 for /i/, (c) log(nasality)–F1 for /u/, (d) log(nasality)–F2 for /a/, (e) log(nasality)–F2 for /i/, (f) log(nasality)–F2 for /u/. The circles, upper triangles and asterisks correspond to O, N and NA, respectively. The solid, dot-dashed and dashed lines are the linear regression fits to the data of O, N and NA, respectively.

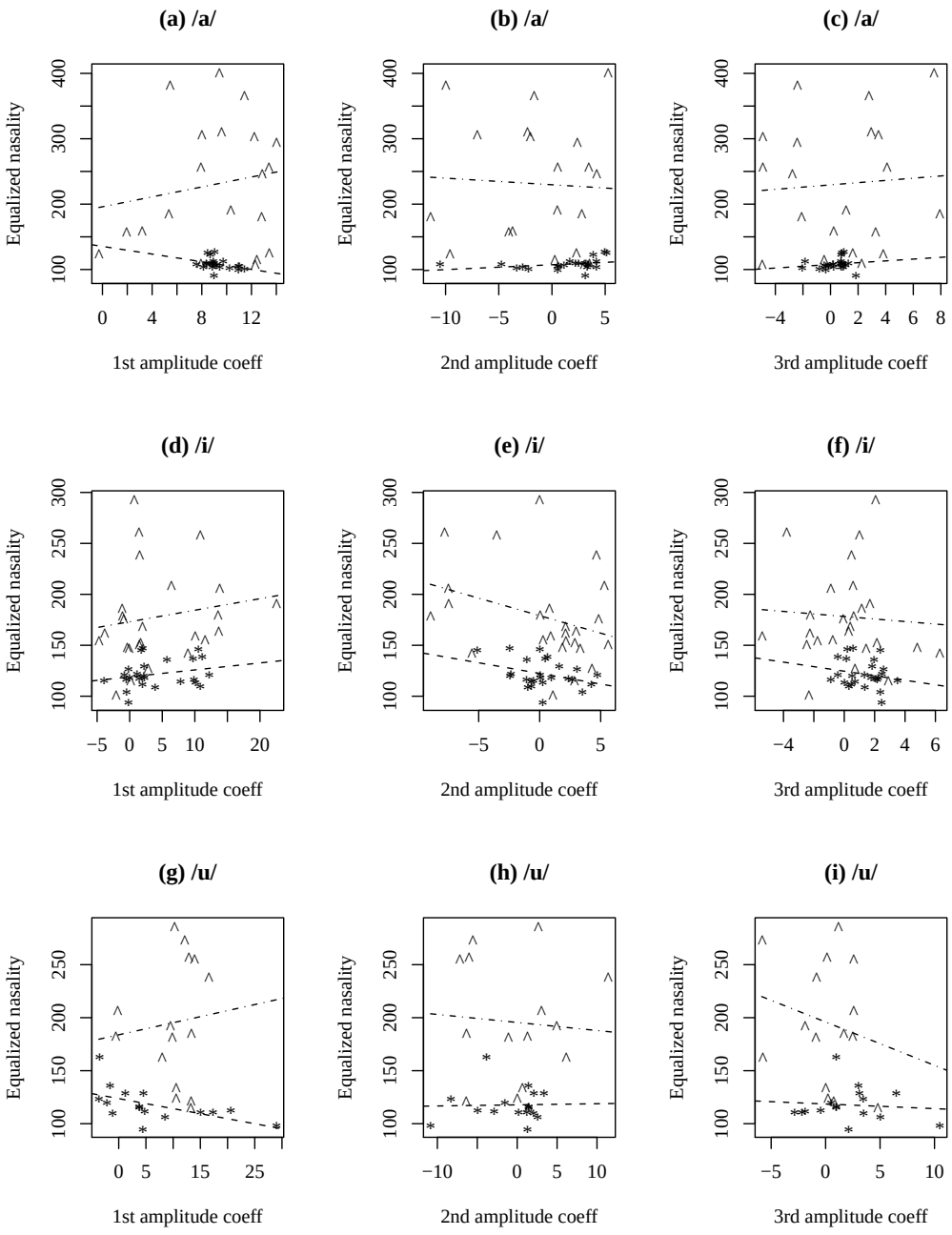


Figure 5.44: Scatterplots of the equalized nasality scores versus the amplitude coefficients of the first three articulatory modes for Speaker 1. (a) nasality–pp1 for /a/, (b) nasality–pp2 for /a/, (c) nasality–pp3 for /a/, (d) nasality–pp1 for /i/, (e) nasality–pp2 for /i/, (f) nasality–pp3 for /i/, (g) nasality–pp1 for /u/, (h) nasality–pp2 for /u/, (i) nasality–pp3 for /u/. The asterisks and upper triangles correspond to NA and N, respectively. The dashed and dot-dashed lines are the linear regression fits to the data of NA and N, respectively.

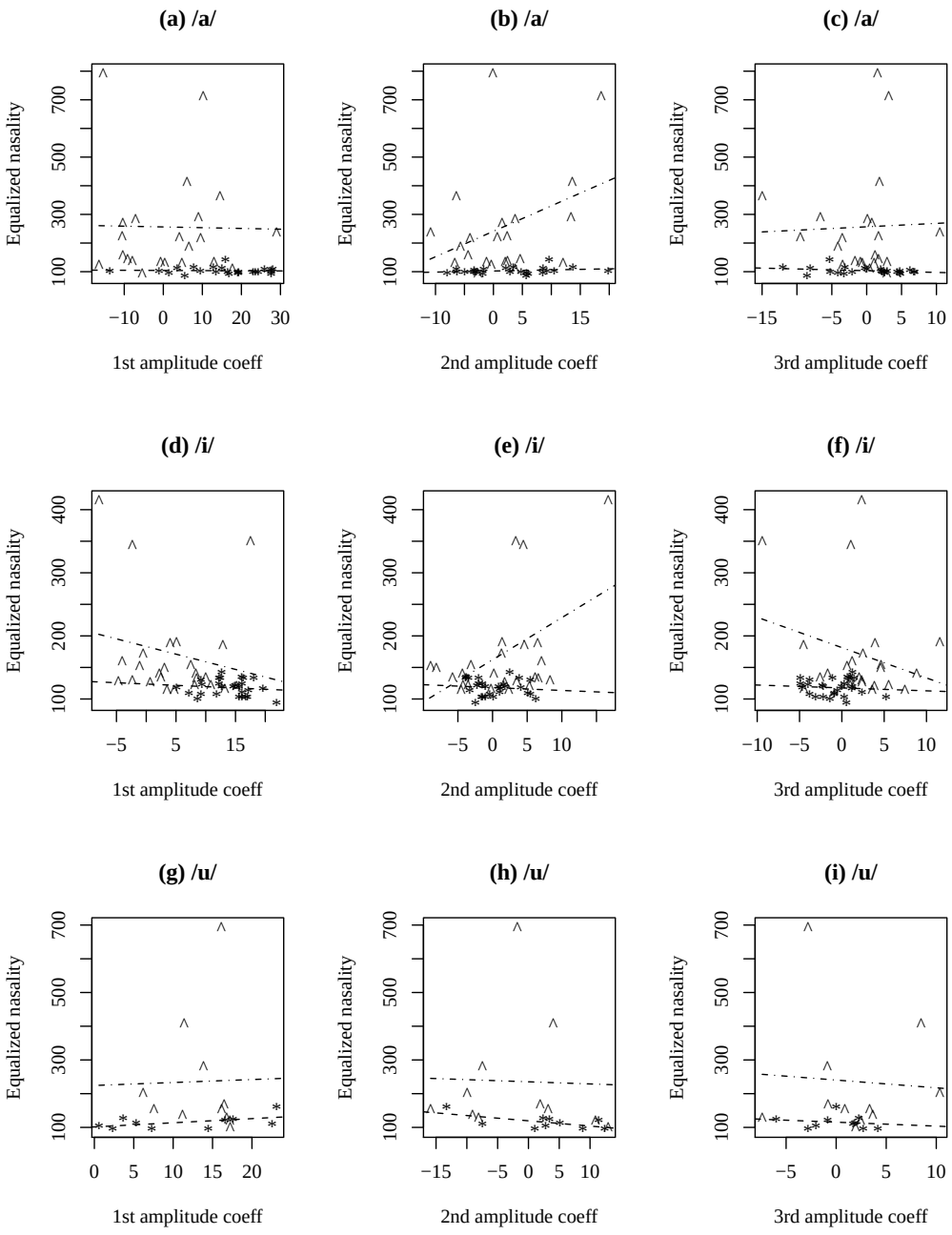


Figure 5.45: Scatterplots of the equalized nasality scores versus the amplitude coefficients of the first three articulatory modes for Speaker 2. (a) nasality–pp1 for /a/, (b) nasality–pp2 for /a/, (c) nasality–pp3 for /a/, (d) nasality–pp1 for /i/, (e) nasality–pp2 for /i/, (f) nasality–pp3 for /i/, (g) nasality–pp1 for /u/, (h) nasality–pp2 for /u/, (i) nasality–pp3 for /u/. The asterisks and upper triangles correspond to NA and N, respectively. The dashed and dot-dashed lines are the linear regression fits to the data of NA and N, respectively.

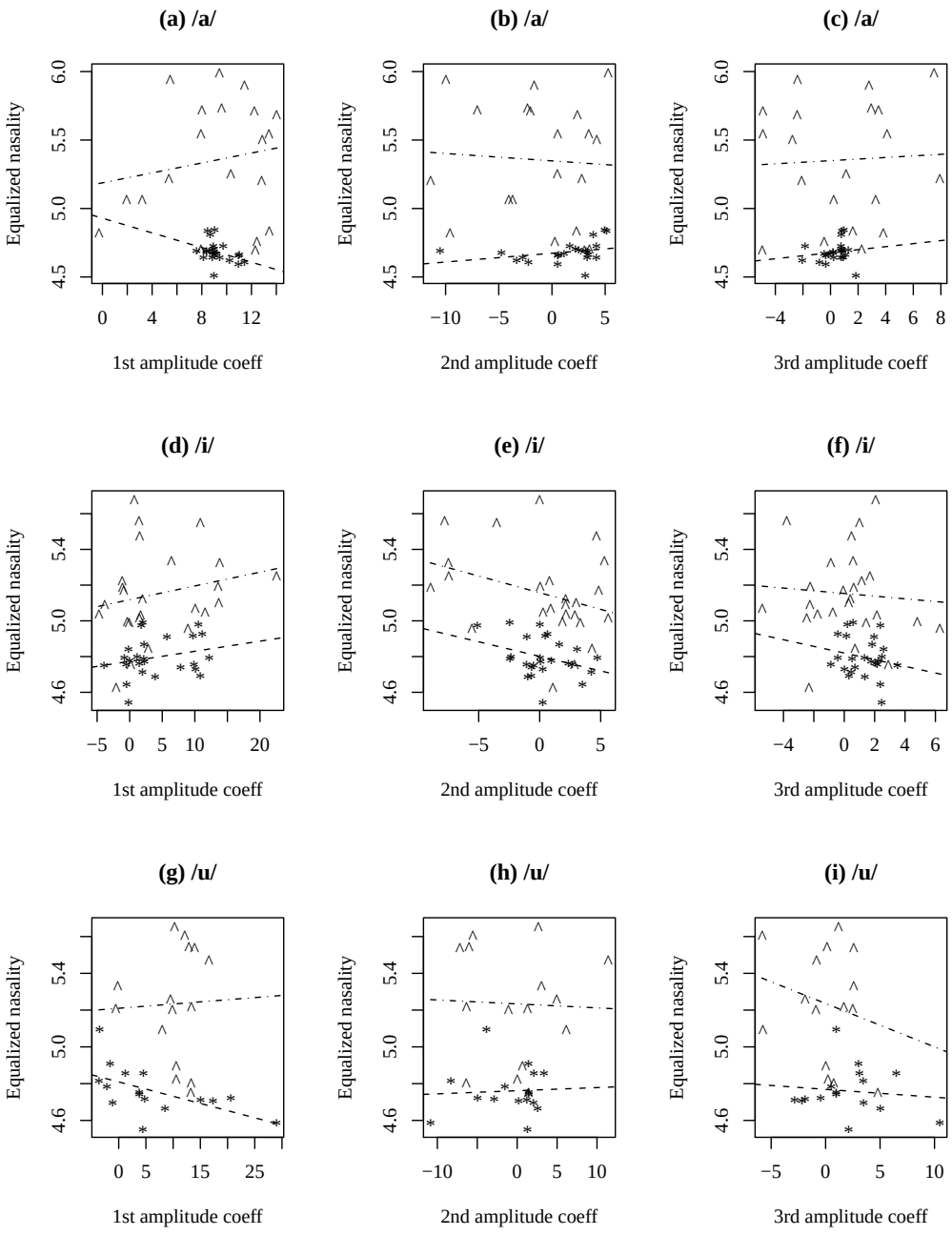


Figure 5.46: Scatterplots of the logarithm of the nasality scores versus the amplitude coefficients of the first three articulatory modes for Speaker 1. (a) nasality–pp1 for /a/, (b) nasality–pp2 for /a/, (c) nasality–pp3 for /a/, (d) nasality–pp1 for /i/, (e) nasality–pp2 for /i/, (f) nasality–pp3 for /i/, (g) nasality–pp1 for /u/, (h) nasality–pp2 for /u/, (i) nasality–pp3 for /u/. The asterisks and upper triangles correspond to NA and N, respectively. The dashed and dot-dashed lines are the linear regression fits to the data of NA and N, respectively.

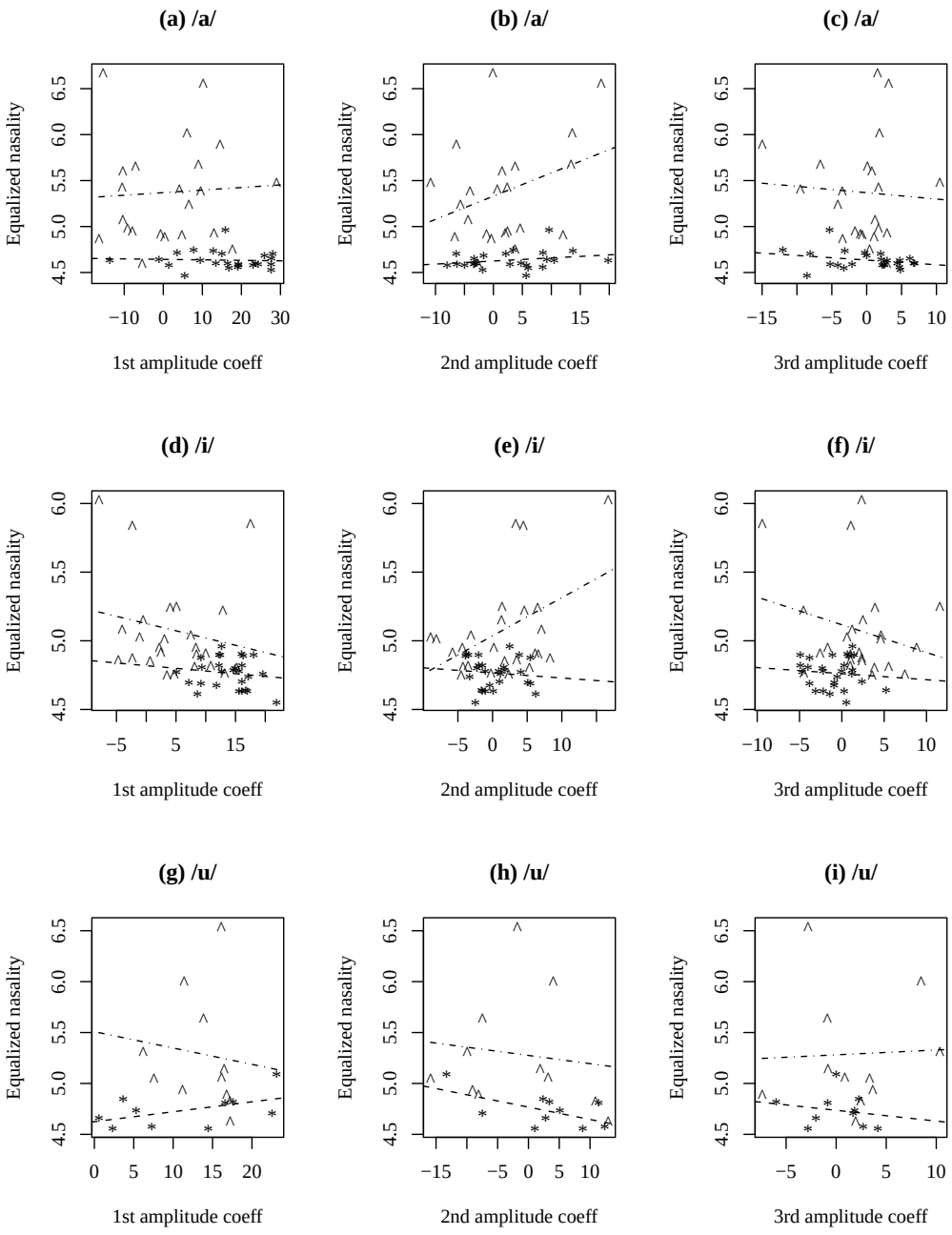


Figure 5.47: Scatterplots of the logarithm of the nasality scores versus the amplitude coefficients of the first three articulatory modes for Speaker 2. (a) nasality–pp1 for /a/, (b) nasality–pp2 for /a/, (c) nasality–pp3 for /a/, (d) nasality–pp1 for /i/, (e) nasality–pp2 for /i/, (f) nasality–pp3 for /i/, (g) nasality–pp1 for /u/, (h) nasality–pp2 for /u/, (i) nasality–pp3 for /u/. The asterisks and upper triangles correspond to NA and N, respectively. The dashed and dot-dashed lines are the linear regression fits to the data of NA and N, respectively.

Appendices

The nasality scores of 54 stimuli were rated twice by each listener on the DME scale. In order to adjust for individual differences of internal perceptual scales and intra-listener variability, the DME nasality scores were equalized following seven procedures Engen (1971):

1. Convert each response to its logarithm;
2. Determine the mean of the logarithm of two responses made by each listener to each stimulus;
3. Determine the mean of step 2, which is regarded as the logarithmic mean of each listener′s responses to all the stimuli;
4. Determine the mean of all the values obtained in step 3 as the logarithmic value of the grand mean of the responses from all listeners to all stimuli;
5. Subtract the value obtained in step 4, the grand mean log response, from each of the individual mean log responses determined in step 3;
6. Subtract the value obtained in step 5 from the mean responses made by each listener to each stimulus in step 2;
7. Convert each value obtained in step 6 to its exponential.

The values obtained in step 7 are regarded as equalized nasality scores.

References

Alwan, A., Narayanan, S., and Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on mri and epg data. part ii. the rhotics. *J. Acoust. Soc. Am.*, 101(2):1078–1089.

Andreassen, M., Smith, B., and Guyette, T. (1991). Pressure-flow measurements for selected oral and nasal sound segments produced by normal adults. *Cleft Palate-Craniofacial Journal*, 28(4):398–407.

Andrews, J. and Rutherford, D. (1973). Contribution of nasally emitted sound to the perception of hypernasality of vowels. *Cleft Palate Journal*, 9:147–156.

Aron, M., Kerrien, E., Berger, M.-O., and Laprie, Y. (2006). Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition set up and preliminary results. *7th International Seminar on Speech Production, Ubatuba, Brazil*, 12.

Badin, P., Beautemps, D., Laboissiere, R., and Schwartz, J. (1995). Recovery of vocal tact geometry from speech signal for vowels and fricative consonants using midsagittal-to-area function conversion model. *Journal of Phonetics*, 23:221–229.

Baer, T., Gore, J., Gracco, L., and Nye, P. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *J. Acoust. Soc. Am.*, 90(2):799–828.

Beautemps, D. and P. Badin, e. a. (1995). Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data. *Speech Communication*, 16(1995):27–47.

Beddor, P. (1983). Phonological and phonetic effects of nasal, on vowel height. Ph.D Dissertation, Indiana University.

Beddor, P. and Hawkins, S. (1990). The influence of spectral prominence on perceived vowel quality. *J. Acoust. Soc. Am.*, 87(6):2684–2704.

Bradford, J., Brooks, A., and Shelton, R. (1964). Clinical judgments of hypernasality in cleft palate children. *Cleft Palate Journal*, 1:329–335.

Bzoch, K. (1989). *Etiological factors related to cleft palate speech. In K. Bzoch (Ed.), Communicative disorders related to cleft palate (3rd).* Little, Brown, Boston.

Carignan, C., Shosted, R., Shih, C., and Rong, P. (2010). Lingual articulation of nasalized vowels in american english. *Journal of Phonetics*, under review.

Carney, P. and Sherman, D. (1971). Severity of nasality in selected speech tasks. *Journal of Speech, Language, and Hearing Research*, 14:396–407.

Chen, M. (1997). Acoustic correlates of english and french nasalized vowels. *J. Acoust. Soc. Am.*, 102(4):2360–2370.

Childers, D. (2000). *Speech Processing and Synthesis Toolboxes.* John Wiley & Sons, Inc., New York, NY.

Coleman, R. (1963). The effect of changes in width in velopharyngeal aperture on acoustic and perceptual properties nasalized vowels. Unpublished doctoral dissertation, Northwestern University.

Colton, R. (1987). The role of pitch in the discrimination of voice quality. *Journal of Voice*, 1(3):240–245.

Counihan, D. and Cullinan, A. (1970). Reliability and dispersion of nasality ratings. *Cleft Palate Journal*, 7:261–270.

Counihan, D. and Cullinan, A. (1972). Some relation- ships between vocal intensity and rated nasality. *Cleft Palate Journal*, 9:101–108.

Dang, J. and Honda, K. (1994). Morphological and acoustical analysis of the nasal and the paranasal cavities. *J. Acoust. Soc. Am.*, 96(4):2088–2100.

Dang, J. and Honda, K. (1996). Acoustic characteristics of the human paranasal sinuses derived from transmission characteristics measurement and morphological observation. *J. Acoust. Soc. Am.*, 100(5):3374–3383.

Delvaux, V. (2003). Controle et connaissance phonetique: Les voyelles nasales du francais. PhD thesis, Universite Libre de Bruxelles, Belgium.

Engen, T. (1971). *Psychophysics II. Scaling Methods. In Woodworth & Schlosbergs Experimental Psychology.* Holt, Rinehart and Winston, Inc.

Engwall, O. (2000a). Are static mri measurements representative of dynamic speech? results from a comparative study using mri, epg and ema. *Proc. ICSLP2000*, I:17–20.

Engwall, O. (2000b). Dynamical aspects of coarticulation in swedish fricatives - a combined ema & epg study. Technical Report 4, KTH TMH-QPSP.

Engwall, O. (2001). Making the tongue model talk: Merging mri and ema measurements. *Proc. Eurospeech*, pages 216–264.

Engwall, O. (2003). Combining mri, ema and epg measurements in a three-dimensional tongue model. *Speech Communication*, 41:303–329.

Engwall, O., Delvaux, V., and Metens, T. (2006). Interspeaker variation in the articulation of nasal vowels. *Proceedings of the 7th ISSP*, page Available online at: http://www.speech.kth.se/prod/publications/files/1926.pdf.

Fant, G. (1960). *The acoustic theory of speech production.* Mouton and Co., The Haugue.

Feng, G. and Castelli, E. (1996). Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization. *J. Acoust. Soc. Am.*, 99(6):3694–3706.

Ferguson, C. (1963). *Assumption about Nasals: A Simple Study in Phonological Universals in Universals of Language.* MIT, Cambridge, MA.

Ferguson, C. (1975). *Universal Tendencies and Normal "Nasality in Nasalfest: Papers from a Symposium on Nasals and Nasalization.* Standard U. P., Standford.

Fletcher, S. (1976). Nasalance vs. listener judgments of nasality. *Cleft Palate Journal*, 13:31–46.

Gay, T., Lindblom, B., and Lubker, J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *J. Acoust. Soc. Am.*, 69(3):802–810.

Gescheider, G. (1976). *Psychophysics, method and theory.* Lawrence Erlbaum, Hillsdale, NJ.

Hardcastle, W., Vaxelaire, B., Gibbon, F., Hoole, P., and Nguyen, N. (1996). Ema/epg study of lingual coarticulation in /kl/ clusters. *Proc. 4th Speech production seminar*, pages 53–56.

Hawkins, S. and Stevens, K. (1985). Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels. *J. Acoust. Soc. Am.*, 77:1560–1575.

Hoole, P., Nguyen-Trong, N., and Hardcastle, W. (1993). A comparative investigation of coarticulation in fricatives: electropalatographic, electromagnetic and acoustic data. *Language and Speech*, 36:235–260.

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

Hughes, O. and Abbs, J. (1976). Labial-mandibular coordination in the production of speech: implications for the operation of motor equivalence. *Phonetica*, 33(3):199–221.

Ito, M., Tsuchida, J., and Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *J. Acoust. Soc. Am.*, 110(2):1141–1149.

Kataoka, R., Warren, D., D.J. Zajaz, R. M., and Lutz, R. (2001). The relationship between spectral characteristics and perceived hypernasality in children. *J. Acoust. Soc. Am.*, 109(5):2181–2189.

Kent, L. (1966). The effect of oral-to-nasal coupling on the perceptual, physiological, and acoustical characteristics of vowels. Unpublished doctoral dissertation, University of Iowa.

Kent, R., Liss, J., and Phillips, B. (1989). *Acoustic analysis of velopharyngeal dysfunction in speech. In K. Bzoch (Ed.), Communicative disorders related to cleft lip and palate (3rd ed).* College-Hill, Boston.

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, 67:971–995.

Maeda, S. (1982a). Acoustic correlates of vowel nasalization: A simulation study. *J. Acoust. Soc. Am.*, 72(Suppl. 1):S102.

Maeda, S. (1982b). The role of the sinus cavities in the production of nasal vowels. In *Proceeding of ICASSP*, pages 991–914.

Maeda, S. (1990). *Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Speech Production and Modeling.* 1990 Kluwer Academic Publishers, Netherlands.

Matsumoto, H., Hiki, S., Sone, T., and Nimuria, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics*, AU21:428–436.

Matsumura, M. (1992). Measurement of 3d shapes of vocal tract and nasal cavity using magnetic resonance imaging technique. In *Proceedings of ICSLP*, pages 779–782, Banff.

McGowan, R. and Cushing, S. (1999). Vocal tract normalization for midsagittal articulatory recovery with analysis-by-synthesis. *J. Acoust. Soc. Am.*, 106(2):1090–1105.

Moen, I. and Simonsen, H. (2007). The combined use of epg and ema in articulatory description. *Advances in Speech-Language Pathology*, 9(1):120–127.

Moore, C. (1992). The correspondence of vocal tract resonance with volumes obtained from magnetic resonance imaging. *Journal of Speech and Hearing Research*, 35:1009–1023.

Moore, W. and Sommers, R. (1973). Phonetic contexts: Their effects on perceived nasality in cleft palate speakers. *Cleft Palate Journal*, 10:72–73.

Naito, M., Deng, L., and Sagisaka, Y. (2003). Model-based speaker normalization methods for speech recognition. *Electronics and Communications in Japan, Part 2*, 86:45–56.

Narayanan, S., Alwan, A., and Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *J. Acoust. Soc. Am.*, 98(3):1325–1347.

Narayanan, S., Alwan, A., and Haker, K. (1997). Towards articulatory-acoustic models for liquid approximants based on mri and epg data, part i. the laterals. *J. Acoust. Soc. Am.*, 101(2):1064–1077.

Nguyen, N., Wrench, A., Gibbon, F., and Hardcastle, W. (1998). Articulatory, acoustic and perceptual aspects of fricative-stop coarticulation. *Proc ICSLP98*, pages 2371–2374.

Overall, J. (1962). Orthogonal factors and uncorrelated factor scores. *Psychological Reports*, 19:651–662.

Perkell, J., Zandipour, M., Matthies, M., and Lane, H. (1999). Articulatory kinematics: preliminary data on the effect of speaking condition, articulator and movement type. *Proc ICPhS99*, 3:1773–1776.

Phillips, B. and Kent, R. (1984). *Acoustic-phonetic descriptions of speech production in speakers with cleft palate and other velopharyngeal disorders. In N. J. Lass (Ed.), Speech and Language: Advances in basic research.* Academic Press, New York.

Pruthi, T., Espy-Wilson, C., and Story, B. (2007). Simulation and analysis of nasalized vowels based on magnetic resonance imaging data. *J. Acoust. Soc. Am.*, 121(6):3858–3873.

Rong, P. and Kuehn, D. (2010). The effect of oral articulation on the acoustic characteristics of nasalized vowels. *J. Acoust. Soc. Am.*, 127(4):2543–2553.

Rong, P. and Kuehn, D. (2012). The effect of articulatory adjustment on reducing hypernasality. *Journal of Speech, Language, and Hearing Research*, doi: 10.1044/1092-4388(2012/11-0142).

Rothenberg, M. (1977). Measurement of air flow in speech. *J. Speech Hear. Res.*, 20:155–176.

Rouco, A. and Recasens, D. (1996). Reliability of electromagnetic midsagittal articulometry and electropalatography data acquired simultaneously. *J. Acoust. Soc. Am.*, 100:3384–3389.

Rubin, P., Baer, T., and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *J. Acoust. Soc. Am.*, 70(2):321–328.

Schiavetti, N., Metz, D., and Silter, R. (1981). Construct validity of direct magnitude estimation and interval scaling: Evidence from a study of the hearing-impaired. *Journal of Speech, Language, and Hearing Research*, 26:568–573.

Schwartz, S. (1968). The acoustics of normal and nasal vowel production. *Cleft Palate Journal*, 9:125–139.

Serrurier, A. and Badin, P. (2008). A 3d articulatory model of the velum and nasopharyngeal wall based on mri and ct data. *J. Acoust. Soc. Am.*, 123(4):2335–2355.

Shosted, R. (2009). *The Aeroacoustics of Nasalized Fricatives: An instrumental study in phonetic typology.* VDM Verlag, Saarbrcken.

Shosted, R., Carignan, C., and Rong, P. (2010). Motor equivalence in the production of phonemic nasal vowels: Evidence from hindi. *J. Acoust. Soc. Am.*, under review.

Singh, S. and Murry, T. (1978). Multidimensional classifica- tion of normal voice qualities. *J. Acoust. Soc. Am.*, 64:81–87.

Stevens, S. (1974). *Handbook of Perception (Vol. 2): Perceptual magnitude and its measurement.* Academic Press, New York.

Stevens, S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects.* Wiley, New York.

Stone, M., Faber, A., Rahael, L., and Shawker, T. (1992). Cross-sectional tongue shape and linguopalatal contact patterns in [s], [ʃ] and [l]. *Journal of Phonetics*, 20:253–270.

Stone, M. and Lundberg, A. (1996). Three-dimensional tongue surface shapes of english consonants and vowels. *J. Acoust. Soc. Am.*, 99:3728–3737.

Story, B. (1995). Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract. Ph.D Thesis, University of Iowa.

Story, B. and Titze, I. (1998). Parameterization of vocal tract area functions by empirical orthogonal modes. *Journal pf Phonetics*, 26:223–260.

Story, B., Titze, I., and Hoffman, E. (1996). Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.*, 100(1):537–554.

Walden, G., Montgomery, A., Gibeily, G., Prosek, R., and Schwartz, D. (1978). Correlates of psychological dimen- sions of in talker similarity. *Journal of Speech, Language, and Hearing Research*, 21:265–275.

Warren, D. (1964). Velopharyngeal orifice size and upper pharyngeal pressure-flow patterns in normal speech. *Plastic reconstr. Surg.*, 33:148–162.

Warren, D. (1967). Nasal emission of air and velopharyngeal function. *Cleft Palate Journal*, 4:148–155.

Warren, D. and Dubois, A. (1964). A pressure-flow technique for measuring velopharyngeal orifice area during continuous speech. *Cleft Palate Journal*, 1:52–71.

Watterson, T., Lewis, K., Allord, M., Sulprizio, S., and O'Neil, P. (2007). Effect of vowel type on reliability of nasality ratings. *Journal of Communication Disorders*, 40:503–512.

Zerling, J. (1984). Phnomnes de nasalit et de nasalization vocaliques: tude cinradiographique pour deux locuteurs. *Travaux de lInstitut de Phontique de Strasbourg*, 16:241–266.

Zhang, Z. and Espy-Wilson, C. (2003). A vocal tract model of american english /l/. *J. Acoust. Soc. Am.*, 115(3):1274–1280.

Zheng, Y., Hasegawa-Johnson, M., and Pizza, S. (2003). Analysis of the three-dimensional tongue shape using a three-index factor analysis model. *J. Acoust. Soc. Am.*, 113(1):478–486.

Zraick, R. and Liss, J. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *Journal of Speech, Language, and Hearing Research*, 43:979–988.

Zraick, R., Liss, J., Case, J., LaPointe, L., and Beals, S. (2000). Multidimensional scaling of nasal voice quality. *Journal of Speech, Language, and Hearing Research*, 43:989–996.