Articulatory differences between oral and nasal vowels based on simulation of a speaker-adaptive articulatory model

Panying Rong¹, Ryan Shosted², David Kuehn³

¹³Department of Speech and Hearing Science, University of Illinois, Champaign, IL, USA
²Department of Linguistics, University of Illinois, Urbana, IL, USA

Abstract

In this study, a speaker-adaptive articulatory model was constructed by fitting point-wise articulatory positions measured by electromagnetic articulography (EMA) to the vocal tract framework in [5] to customize a standard vocal tract model with speaker-dependent articulatory features. With the speaker-adaptive articulatory model, the area functions of oral and nasal vowel pairs (/a/, /i/, /u/) were simulated. The differences of area functions between oral and nasal vowels were decomposed into orthogonal modes [10] to account for the primary oropharyngeal articulatory changes related to nasalization. The relationship between the principal articulatory modes and the lowest two formant frequencies (F1, F2) was examined to uncover the effect of oropharyngeal articulation on the acoustics of nasal vowels.

Index Terms: articulatory modeling, nasalization, electromagnetic articulography

1. Introduction

Under the common assumption that the complex activities of speech articulators can be decomposed and represented by the coordinated activities of a limited set of independent functional units, articulatory models are usually constructed by 1) partitioning the entire vocal tract into a number of independent functional units; 2) controlling the positions of the functional units with a set of discrete parameters; and 3) coordinating the movements of the functional units to achieve an acoustic target through inverse filtering. Such traditional articulatory models [5] rely on a standard anatomical framework and fixed articulatory dimensions of the vocal tract, which cannot account for the significant inter-speaker variability associated with divergent anatomical and articulatory characteristics.

To overcome the inconsistency between the traditional model and a speaker-dependent vocal tract, a speaker-adaptive normalization is needed to customize the standard vocal tract model with speaker-dependent features that account for inter-speaker anatomical variability (e.g., size and shape of anatomic structures) and articulatory flexibility.

[6] proposed a two-step speaker normalization for the Haskins Laboratories articulatory synthesizer, ASY. The length and transverse dimensions of the ASY vocal tract model were first adapted to individual speakers; then, using computation based on artificial neural networks, the midsagittal articulatory configuration of the model was transformed to fit the articulatory positions of individual speakers (measured by X-ray microbeam).

The second step is an articulatory normalization using an inverse normalizing map (NM). According to [6], NM is the transformation of midsagittal shapes or articulatory positions of a speaker-dependent vocal tract to a standard vocal tract. [6] also defined the parameters of the standard vocal tract that are not affected by NM, such as the length and transverse dimensions of the vocal tract, as PONMs (Parameters Orthogonal to the space of NM). Thus, the first step is actually a normalization of the PONMs, which is intended to match the acoustic outputs of the model and the speaker mapped to one another under the NM or its inverse [1, 2].

In this study, we constructed a speaker-adaptive articulatory model by adapting the point-wise EMA articulatory measurements to the framework of the standard vocal tract model in [5]. On the basis of [6], a normalization procedure was developed to first adjust the PONMs of the standard model and then adapt the articulatory configuration of the model to the EMA articulatory measurements of a speaker-dependent vocal tract. With the speaker-adaptive articulatory model, we simulated the area functions of oral and nasal vowels /a/, /i/, and /u/, and examined the interaction of nasalization and oropharyngeal articulation of nasal vowels. Previous studies have compared the differences of individual articulatory positions between oral and nasal vowels [3, 9, 4]. However, we know of no study that directly compares oropharyngeal articulatory differences between oral and nasal vowels with the acoustic outcomes of those differences. The current study will use the area functions of oral and nasal vowels (estimated from the articulatory model) to derive the principal articulatory modes responsible for acoustic differences observed between nasal and oral vowels (other than those associated with velopharyngeal opening itself).

2. Method

2.1. Participant and speech materials

The participant was a native speaker of American English with no reported speech pathologies. Elicitation items consisted of /CV/, /VC/, /CV/C/ and /VCV/ syllables, where V includes the three corner vowels /a/, /i/ and /u/; and C includes stop, fricative and nasal consonants with different places of articulation (i.e., /p/, /b/, /t/, /d/, /k/, /g/, /f/, /v/, /s/, /z/, /m/, /n/, /l/). Different consonantal contexts varying in place and manner of articulation were included to account for the coarticulatory effect of consonantal articulation on vowel articulation. Each syllable was recorded twice within one of the following carrier phases: “Say ... again,” “Say ... six times,” “I said ... again” and “I
said ... six times,” depending on the structure of the syllable. A total of 264 stimuli were recorded and the vowel portions of all syllables were segmented to compose the database for speaker adaptation of the articulatory model.

2.2. Database acquisition

The database for model adaptation consisted of simultaneously-recorded articulatory, acoustic, and nasal-aerodynamic signals of the speaker. The instrumentation for data acquisition followed the same setup as [3, 9].

2.2.1. Articulatory and acoustic data acquisition

Articulatory data were collected with the Carstens AG500 EMA system at a sampling rate of 200Hz. Twelve electromagnetic sensors were attached to the face and articulators of the speaker using surgical glue to track their movements. Among the 12 sensors, nine were attached to the articulators of the subject: four sensors on the tongue, including one on the tongue tip (TT), another on the tongue blade (TB: 0.5–1.0cm posterior to TT), and two on the tongue dorsum (PTD: 4.0cm posterior to TB and ATD: midway between TB and PTD); two on the teeth (UL/LL: upper and lower incisors); two on the lips (UL/LL: upper and lower borders of vermilion in midline); and one on the chin. Three additional sensors were affixed to the bony structures (nose bridge, left and right tragi) to serve as reference sensors for head movement correction. A head-mounted microphone was positioned about 5cm away from the subject’s left lip corner to record the audio signal simultaneously with articulatory movements.

2.2.2. Nasal aerodynamic data acquisition

Measurement of nasal pressure/flow gives an indirect cue of velar movement. The onset/offset of nasal airflow can indicate the onset/offset of vowel nasalization across tokens for a given speaker [8, 12]. To record nasal airflow, the participant wore a vented Scicon NM-2 nasal mask [7]. The air pressure signal collected by the nasal mask was transformed by a Biopac TSD 160A pressure transducer, digitized at 1 kHz, and recorded using custom-written scripts running in Matlab 7.5.0 (R2007b).

2.2.3. System synchronization

The EMA, acoustic and aerodynamic signals were synchronized by the Sybox-Opto4 included with the AG500 system.

2.2.4. Recording

Prior to speech recording, the subject was asked to talk at a conversational level with the EMA sensors and the nasal mask attached to get accustomed to the experimental setup. The recording started when the subject felt comfortable with the experimental setup and his speech was perceived to be natural by the experimenters. During the recording, the subject was asked to read the speech stimuli displayed on a series of PowerPoint slides on a laptop monitor controlled by one of the experimenters. Another experimenter monitored the real-time display of articulatory movements using CS5View (Carstens). At the end of recording, an experimenter traced the midsagittal contour of the subject’s hard palate using one of the sensors.

2.3. Data annotation

For each token, three acoustic landmarks were annotated: the onset, midpoint, and offset of the vowel. Vowel onset/offset was determined as the onset/offset of regular vibration in the sound pressure waveform. For nasal tokens, an additional landmark characterizing the onset or offset of vowel nasalization was annotated. With a threshold at 20% above the average of filtered nasal airflow velocity, the first positive nasal airflow velocity peak above the threshold that occurred after the voice onset of the vowel and the last positive velocity peak above the threshold that occurred before the offset of the vowel were determined to be the onset and offset of nasalization, respectively. In addition to the acoustic landmarks, the following articulatory landmarks related to the movement of the primary articulator were also annotated: 1) the maximum vertical position of TT for /l/, /l/ and /l/; /l/; /l/; /l/; /l/; 2) the maximum position of PTD for /k/, /g/, /g/; and 3) the maximum position of the UL and the minimum position of LL for /p/, /b/, /b/.

2.4. Articulatory modeling

The speaker-adaptive articulatory model was constructed following a six-step paradigm: 1) fitting the EMA articulatory positions to the framework of the vocal tract model in [5]; 2) deriving the midsagittal shape of the vocal tract; 3) estimating the constraints on articulatory movement; 4) estimating the formant frequencies of all vowel samples based on a transmission line model of the vocal tract; 5) optimizing the PONMs, including the length and the transverse dimensions of the model; and 6) computing the area functions of the vocal tract.

First of all, the midsagittal palatal trace measured by EMA was fitted to the hard palate of the model. The shape of the soft palate was determined by the velopharyngeal opening area, which was assumed to be zero for oral vowels and 220mm² for nasal vowels. Using a physiology-based interpolation [5], the midsagittal configuration of the vocal tract was estimated given nine discrete articulatory positions: the configuration of the upper pharyngeal cavity was determined by PTD; the configuration of the posterior oral cavity was determined by PTD and ATD; the anterior oral cavity was determined by TB, TT, and LL; and lip opening and protrusion were determined by UL and LL. Due to a lack of EMA measurements for the lower pharynx, our model used the default shapes for /a/, /i/, and /u/ given in [5] to represent the corresponding lower pharyngeal configurations. The constraints on articulatory movement were then estimated from the ranges of the articulatory positions across all speech samples in the database. Based on transmission line theory, the lowest three formant frequencies (i.e., F1, F2, F3) of all vowels samples were computed from the estimated midsagittal configuration and the default PONMs (i.e., length and transverse dimensions) of the vocal tract model. Using a simulated annealing algorithm [11], the PONMs of the model were optimized by matching the formant frequencies of the model and the corresponding acoustic signal. Finally, the area function was calculated based on the midsagittal configuration and the optimized PONMs of the vocal tract.

2.5. Orthogonal articulatory mode decomposition

To compare the oropharyngeal articulation of oral and nasal vowels, the estimated area functions of the oral and nasal vowels...
were used to decompose a series of orthogonal modes responsible for the primary oropharyngeal articulatory differences between the oral and nasal vowels. The decomposition followed procedures similar to those in [10], and can be represented by the following equations:

\[
\alpha(v, s) = A(v, s) - A_0(v)
\]

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

\[
R_{ij} = \frac{1}{M-1} \sum_{s=1}^{M} \alpha(v_i, s) \alpha(v_j, s)
\]

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

\[
R\phi = \phi I \lambda \tag{3}
\]

where \( R \) is the covariance matrix in Eq. 2, \( I \) is the identity matrix, \( \phi \) and \( \lambda \) are the empirical orthogonal modes and their corresponding eigenvalues, respectively.

\[
pp(i, s) = \sum_{j=1}^{N} \alpha(v_j, s) \phi(i, j)
\]

where \( i = 1, 2, ..., N \) corresponding to the \( i^{th} \) mode and \( pp(i, s) \) is the amplitude coefficient obtained by the projection of the \( s^{th} \) \( \alpha \) vector on the \( i^{th} \) mode.

3. Results

Figure 1 gives an example of the mid sagittal vocal tract configuration for an oral /i/ sample estimated from EMA measurements, which are represented by various markers in the figure.

\[\begin{align*}
\text{Figure 1: Mid sagittal configuration of the vocal tract estimated for the initial /i/ in /ipʰ/. The four markers on the tongue correspond to PTD (circle), ATD (asterisk), UB (triangle), and TT (dot). The asterisk and dot markers at the opening of the vocal tract correspond to UL and LL respectively. The asterisk in the hard palate region represents UI and the dot behind LL corresponds to LI. The dot below LL marks the position of the chin. The upper triangle and two circles above the palate correspond to the right and left tragi and the nose, respectively. In addition to the EMA measurements, those critical articulatory positions used in construction of the model are also marked in the figure, including the center of each part of the tongue (two asterisks, one circle and one dot) and the hyoid bone position (bottom triangle).}
\end{align*}\]

The relationships between the amplitude coefficients (pp1, pp2, pp3) and the lowest two formant frequencies (F1, F2) of the nasal vowels were plotted in Figures 3–5, where pp1, pp2, and pp3 are the amplitude coefficients computed as the projections of the \( \alpha \) vector on the first three modes, respectively.

The following correlations between amplitude coefficients and formant frequencies were found to be significant: 1) pp3–F1 for /a/ \(( R = -0.532, p = .009)\); 2) pp1–F2 for /i/ \(( R = .704, p = .002)\); 3) pp2–F2 for /i/ \(( R = -0.426, p = .043)\); 4) pp1–F1 for /u/ \(( R = -0.417, p = .048)\); 5) pp1–F2 for /u/ \(( R = -0.599, p = .005)\); 6) pp2–F2 for /u/ \(( R = .632, p = .003)\).

Comparisons of formant frequencies between oral and nasal vowels showed the following significant differences: 1) /i/: \( F1_{oral} < F1_{nasal} \ (F(1, 87) = 11.58, p = .001)\); 2) /i/: \( F2_{oral} > F2_{nasal} \ (F(1, 87) = 31.97, p < .001)\); 3) /u/: \( F1_{oral} < F1_{nasal} \ (F(1, 62) = 24.23, p < .001)\); 4) /a/: \( F2_{oral} > F2_{nasal} \ (F(1, 62) = 23.09, p < .001)\).

4. Conclusions

A speaker-adaptation was applied to adjust the PONMs and articulatory dimensions of a standard vocal tract model [5] to match a speaker-dependent vocal tract through fitting of point-wise articulatory positions measured using EMA. With the area functions of oral and nasal vowels estimated from the speaker-adaptive articulatory model, a set of empirical orthogonal modes were decomposed from the area functions to indicate the oropharyngeal articulatory differences between oral and nasal vowels. The oral-nasal articulatory differences were then correlated with formant frequencies to indicate the effect of oropharyngeal articulation on the acoustics of nasal vowels.

As the F1 frequency of nasal /i/ is significantly lower than its oral counterpart and a lower F1 frequency is correlated with a higher amplitude coefficient pp1 (Figure 4), it suggests that the articulatory gestures represented by the first orthogonal mode for /i/ in Figure 2 are responsible for the F1 difference between oral and nasal /i/. Such articulatory gestures include expansions in the front oral cavity to accompany lowering of the velum (Figure 2). Recruiting such articulatory gestures during nasalization also leads to higher F2 frequencies, although the correlation between F2 and pp1 is not significant (Figure 4). The F2 frequency of nasal /a/ is significantly higher than its oral counterpart and a higher F2 is correlated with both higher pp1 and pp2 (Figure 5), so the articulatory gestures represented by the first and second orthogonal modes for /a/ in Figure 2 are responsible for the F2 difference between oral and nasal /a/. Such articulatory gestures consist of a constriction in the front oral cavity (represented by the first and second modes) and an expansion in the upper pharynx (represented by the second mode). Recruiting such articulatory gestures also leads to significantly lower F1 frequencies (Figure 5) for nasal /a/ compared to oral /a/. For /u/, as the first and second orthogonal modes have opposite effects on the formant frequencies (i.e., a higher pp1 is related to lower F1 and higher F2, whereas a higher pp2 is related to higher F1 and lower F2) (Figure 3), the oral-nasal differences of F1 and F2 are found to be insignificant.
Figure 2: The first three orthogonal modes decomposed from the area differences between nasal and oral vowels (solid curves). From top to bottom, the three rows correspond to /a/, /i/ and /u/, respectively. From left to right, the three columns correspond to the first, second and third orthogonal modes, respectively.

Figure 3: Relationships between amplitude coefficients pp1, pp2, pp3 and F1 for nasal /a/ from left to right in the first row. Relationships between pp1, pp2, pp3 and F2 for nasal /a/ from left to right in the second row. The straight lines are linear regression fits to the data.

Figure 4: Relationships between amplitude coefficients pp1, pp2, pp3 and F1 for nasal /i/ from left to right in the first row. Relationships between pp1, pp2, pp3 and F2 for nasal /i/ from left to right in the second row. The straight lines are linear regression fits to the data.

Figure 5: Relationships between amplitude coefficients pp1, pp2, pp3 and F1 for nasal /u/ from left to right in the first row. Relationships between pp1, pp2, pp3 and F2 for nasal /u/ from left to right in the second row. The straight lines are linear regression fits to the data.

Therefore, the speaker-adaptive articulatory model provides an effective way to simulate speaker-dependent articulatory configurations when structural images are not available. This may have potential applications in speech pathology, e.g., to assist in the diagnosis and treatment of articulatory disturbances. Based on the simulation of area functions, the oral-nasal articulatory differences (other than differences in velopharyngeal opening) responsible for the acoustic discrepancies (F1, F2) between oral and nasal vowels were derived as empirical orthogonal modes to help uncover the articulatory-acoustic relationship in nasal vowels.

5. References


