Vocal tract modes based on multiple area function sets from one speaker
Brad H. Story

Citation: The Journal of the Acoustical Society of America 125, EL141 (2009); doi: 10.1121/1.3082263
View online: https://doi.org/10.1121/1.3082263
View Table of Contents: https://asa.scitation.org/toc/jas/125/4
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

Vocal tract area functions from magnetic resonance imaging
The Journal of the Acoustical Society of America 100, 537 (1996); https://doi.org/10.1121/1.415960

A parametric model of the vocal tract area function for vowel and consonant simulation
The Journal of the Acoustical Society of America 117, 3231 (2005); https://doi.org/10.1121/1.1869752

Identification of synthetic vowels based on a time-varying model of the vocal tract area function
The Journal of the Acoustical Society of America 127, EL146 (2010); https://doi.org/10.1121/1.3313921

Vowel and consonant contributions to vocal tract shape
The Journal of the Acoustical Society of America 126, 825 (2009); https://doi.org/10.1121/1.3158816

A portable infrasound generator
The Journal of the Acoustical Society of America 125, EL148 (2009); https://doi.org/10.1121/1.3093797

An age-dependent vocal tract model for males and females based on anatomic measurements
The Journal of the Acoustical Society of America 143, 3079 (2018); https://doi.org/10.1121/1.5038264
Vocal tract modes based on multiple area function sets from one speaker

Brad H. Story
Speech Acoustics Laboratory, Department of Speech, Language, and Hearing Sciences,
University of Arizona, Tucson, Arizona 85721
bstory@u.arizona.edu

Abstract: The purpose of this study was to derive vocal tract modes from a wider range of vowel area functions for a specific speaker than has been previously reported. Area functions from Story et al. [(1996). J. Acoust. Soc. Am. 100, 537–554] and Story [(2008). J. Acoust. Soc. Am. 123, 327–335] were combined in a composite set from which modes were derived with principal component analysis. Along with scaling coefficients, these modes were used to generate a [F1, F2] formant space. In comparison to formant spaces similarly generated based on the two area function sets alone, the combined version provides a wider range of both F1 and F2 values. This new set of modes may be useful for inverse mapping of formant frequencies to area functions or for modeling of vocal tract shape changes.

© 2009 Acoustical Society of America
PACS numbers: 43.70.Bk, 43.70.Aj [AL]
Date Received: September 7, 2008 Date Accepted: January 8, 2009

1. Introduction

For production of vowels, the vocal tract area function has been shown to be fairly well represented by two canonical deformation patterns, or “modes,” (e.g., Story and Titze, 1998; Story, 2005b; Mokhtari et al., 2007) derived from principal component analysis (PCA). Such modes have been derived from speaker-specific area function sets, but are similarly shaped (in terms of their variation along the length of the vocal tract) across speakers, and are related to specific patterns of formant frequencies when appropriately scaled and superimposed on a mean area function.

The modal representation of the vocal tract shape allows for an essentially one-to-one mapping between the scaling coefficients of the modes and the first two formant frequencies and has led to a method for mapping time-varying formant frequencies extracted from recorded speech to a time-varying sequence of area functions (Story and Titze, 1998; Story and Titze, 2002; Mokhtari et al., 2007). Development of a kinematic model of the vocal tract area function that can be used to simulate speech is also based on this same modal representation (Story, 2005a). Although both the inverse mapping technique and kinematic model have been shown to be reasonably successful at bridging the area function-to-acoustic relation, the ability of the modal representation to produce a wide range of vocal tract shapes is limited by the original set of area functions on which it is based. For example, the modal representation derived in Story and Titze (1998) and subsequently utilized in Story and Titze (2002) and Story (2004, 2005b) was based on the ten vowel area functions reported for an adult male (Story et al., 1996). Hence, the boundaries of the potential [F1, F2] vowel space afforded by the modal representation are essentially defined by the formant frequencies of those original area functions. Although additional sets of speaker-specific area functions reported by Story (2005b) and Mokhtari et al. (2007) have been used to derive modal representations of individual speakers, they are also limited by the range of formant frequencies produced by the original area functions obtained from each speaker. This is not necessarily a severe limitation assuming that each speaker produced hyperarticulated vowels (thus, producing extreme [F1, F2] formant values). If the vowels tend toward being centralized, however, the working space of a modal-based vocal tract model will be constrained.
Recently, Story (2008) reported an additional set of 11 vowel area functions for the same speaker who produced the 10 vowels originally reported in Story et al. (1996). F1 and F2 formant values were determined from the frequency response function calculated for each area function in the two sets. It was noted that, in general, the F2 values calculated for the new area function set were shifted downward in frequency relative to those calculated for the area functions of Story et al. (1996) as shown in Fig. 1. It was suggested that this difference originated from a tendency, in the new area functions, to slightly constrict the pharyngeal portion of the vocal tract while expanding the oral cavity. Since the new area functions are configured somewhat differently than the 1996 versions and produce different [F1, F2] values, they provide additional samples of the same speaker’s vocal tract that might be combined with the original set in a PCA. The resulting vocal tract modes and mean area function would then be representative of a wider range of vowel configurations than either area function set alone.

The purpose of this letter is to demonstrate that a modal representation derived from a particular combination of area functions selected from the data of Story et al. (1996) and Story (2008) produces an [F1, F2] vowel space that is larger than that produced by modes derived from either area function set alone.

2. Principal component analysis

Area function sets from both Story et al. (1996) and Story (2008) were used in the analysis. The area functions of Story et al. (1996) included the vowels /i i e æ a o o u/ and will henceforth be referred to as “A96.” In addition to these same ten vowels, the Story (2008) data include the vowel /e/; this set of 11 area functions will be referred to as “A08.”

Any given area function can be represented by two components, an area vector and a length increment. The area vector $A(i)$ contains 44 cross-sectional areas, assumed to represent a concatenation of tubelets ordered consecutively from glottis to lips. The index $i$ denotes this ordering and extends from 1 to 44. The length increment $\Delta$ is the distance between consecutive cross-sectional areas and can be considered to be the tubelet length. The 44-section area vectors and length increments, as used in the present study, were reported in Story and Titze (1998) for A96 and in Story (2008) for A08. The A96 set, however, was smoothed for this study in the same manner as the A08 set (see Story, 2008, p. 328).

As mentioned in the Introduction, the calculated [F1, F2] formant frequencies for each of the two area function sets are shown in Fig. 1. The open circles indicate the vowels in the plot...
that are representative of the boundaries of the displayed vowel space, regardless of set membership, and their corresponding area functions served as the combined data set on which a new PCA was performed. This combined set is referred to as “A9608x” and specifically includes [i e æ u] from A96 and [æ u o o u] from A08, where the “x” denotes that the combined set excludes some of the vowels in A96 and A08. Note that the target vowels [æ e] and [u] from both A96 and A08 are included in the combined set because they represent part of the overall boundary of the vowel space.

The collection of area vectors for a given set (i.e., A96, A08, or A9608x) can be represented in matrix form as \( A(i,j) \), where \( i \) is the area index and \( j \) denotes the particular vowel in the clockwise order indicated by the points along the dashed or solid lines for A96 and A08, respectively, or by the large circles in Fig. 1 for A9608x (i.e., \( j = 1 \) for [i], \( j = 2 \) for [e]...\( j = 11 \) for [u]). Following Story (2005b), the PCA was performed on the equivalent diameters of the cross-sectional areas rather than on the areas themselves. In addition, the length of each area function was included in the same manner as Yehia et al. (1996) and Mokhtari et al. (2007) where the variance of the length increment \( \Delta \) is first normalized by the largest variance of the equivalent diameters, and then becomes the 45th element. Thus, a matrix \( D(i,j) \), containing直径 and length information, is constructed as

\[
D(i,j) = \begin{cases} 
\sqrt{\frac{4}{\pi} A(i,j)} & \text{for } i = 1, \ldots, 44 \\
\frac{(\Delta_j - \bar{\Delta})\sigma_{\Delta}}{\sigma_D} + \bar{\Delta} & \text{for } i = 45, 
\end{cases}
\]

where \( \Delta_j \) is the length increment of area vector \( j \), \( \bar{\Delta} \) is the mean length increment, \( \sigma_{\Delta} \) is the standard deviation of the length increments, and \( \sigma_D \) is the largest standard deviation within any section \( i \) of the equivalent diameters.

Matrix \( D(i,j) \) can be represented by a mean and variable part,

\[
D(i,j) = \Omega(i) + \alpha(i,j),
\]

where \( \Omega(i) \) is the mean vector across \( D(i,j) \), and \( \alpha(i,j) \) is the variation superimposed on \( \Omega(i) \) to produce a specific vector. The PCA was then carried out by calculating the eigenvectors of a covariance matrix formed with \( \alpha(i,j) \). The specific implementation was essentially the same method as reported in Story and Titze (1998), and results in the following parametric representation of the original \( D(i,j) \) matrix:

\[
\hat{D}(i,j) = \Omega(i) + \sum_{j=1}^{N} q(j) \phi_l(i), \quad i = [1,N] \quad (N = 45), \quad j = [1,11],
\]

where the \( \phi_l(i) \)'s are 45-element eigenvectors (modes) that, when multiplied by the appropriate scaling coefficients \( q(j) \), will reconstruct each area vector in \( A(i,j) \) for \( i = [1,44] \) and each length increment for \( i = 45 \) by the following:

\[
\hat{A}(i,j) = \frac{\pi}{4} \hat{D}(i,j)^2, \quad i = [1,44], \quad j = [1,11],
\]

\[
\hat{\Delta}(j) = \frac{(\hat{D}(45,j) - \bar{\Delta})\sigma_{\Delta}}{\sigma_D} + \bar{\Delta}, \quad j = [1,11].
\]

Area functions generated from the mean diameter function \( \Omega(i) \) [i.e., \( q(j) = 0 \)] of each set A96, A08, and A9608x are plotted against the normalized distance from the glottis in Fig. 2(a). It is noted that the area function based on the combined set (thick line) is more constricted in the
pharyngeal portion (from approximately 0.2 to 0.6 along the length) and more expanded in the oral portion (from 0.6 to 1) than either of the other two.

The two modes that accounted for most of the variance in the analysis of the three area function sets [referred to as \(\phi_1(i)\) and \(\phi_2(i)\)] are shown in Fig. 2(b), again plotted against the normalized distance from the glottis. Note that the modes have been smoothed by fitting them with eighth-order polynomials. This simplifies the visual comparison of the modes but maintains their gross characteristics. As in Fig. 2(a), the thick, thin, and dashed lines represent the A9608x, A96, and A08 sets, respectively. The variances accounted for by each mode in each of the three sets are given in Table 1. The first mode \(\phi_1\) accounts for greater than 64% of the variance, whereas \(\phi_2\) accounts for about 20%. The total amount of the variance accounted for by these two modes was greatest for the A9608x set with 89.6%. Qualitatively there are some minor differences in the amplitude of the modes across the three sets, but the overall shape is essentially the same. That is, with a positive scaling coefficient, \(\phi_1\) will expand the front half of the vocal tract and constrict the back half, and vice-versa with a negative scaling coefficient. When positively scaled, the second mode imposes an expansion near the lips, followed by a constriction, an expansion, and another constriction above the glottis; the opposite effect occurs with a negative scaling coefficient.

3. Vowel space calculations

A mapping was generated for each area function set that relates the two scaling coefficients \(q_1\) and \(q_2\) to the first two formant frequencies (F1 and F2). Based on the PCA performed on each

<table>
<thead>
<tr>
<th>Mode</th>
<th>A96</th>
<th>A08</th>
<th>A9608x</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\phi_1)</td>
<td>64.8</td>
<td>69.8</td>
<td>65.9</td>
</tr>
<tr>
<td>(\phi_2)</td>
<td>21.3</td>
<td>18.5</td>
<td>23.7</td>
</tr>
<tr>
<td>Total</td>
<td>86.1</td>
<td>88.3</td>
<td>89.6</td>
</tr>
</tbody>
</table>
set, an equal increment continuum was generated for each mode coefficient that ranged from their respective minimum to maximum values. The increments were specified as

\[ \Delta q_1 = \frac{q_1^{\text{max}} - q_1^{\text{min}}}{M-1}, \]

\[ \Delta q_2 = \frac{q_2^{\text{max}} - q_2^{\text{min}}}{N-1}, \]

where \( M \) and \( N \) are the numbers of increments along each coefficient dimension. For this study \( M = N = 80 \). The coefficient continua were then generated by

\[ q_{1m} = q_1^{\text{min}} + m\Delta q_1, \quad m = 0, \ldots, M-1, \]

\[ q_{2n} = q_2^{\text{min}} + n\Delta q_2, \quad n = 0, \ldots, N-1. \]

with \( m \) and \( n \) serving as indices along each continuum. By modifying Eq. (3) to contain only two modes and eliminating the dependence on a specific vowel (i.e., \( j \)th vowel), an area vector and length increment can be directly generated with

\[ A_{mn}(i) = \frac{\pi}{4} \left[ \Omega(i) + q_{1m}\phi_1(i) + q_{2n}\phi_2(i) \right]^2, \quad i = [1, 44], \]

\[ \Delta_{mn} = \frac{(\Omega(i) + q_{1m}\phi_1(i) + q_{2n}\phi_2(i) - \bar{\Delta})\sigma_A}{\sigma_D} + \bar{\Delta}, \quad i = 45. \]

Shown in Fig. 3(a) is the \( 80 \times 80 \) coefficient grid for the A9608x set along with the coefficient pairs that reconstruct the original 11 vowels used in the PCA for this set. The squares and circles represent those vowels extracted from the A96 and A08 sets, respectively, to make up the combined set. Also shown in this figure are the outlines of coefficient grids based on the separate
PCAs of the A96 (solid line) and A08 (dashed line) sets. The coefficient ranges are similar for the A9608x and A08 sets, but the range for the A96 set is contracted relative to the other two. It is noted that all three sets have nearly the same upper limit of the $q_2$ coefficient.

For each of the 6400 coefficient pairs in the grid (i.e., every intersection point), an area vector and length increment were generated. A frequency response function of each resulting area function was then calculated with a frequency-domain technique based on cascaded “ABCD” matrices (Sondhi and Schroeter, 1987; Story et al., 2000). This calculation included energy losses due to yielding walls, viscosity, heat conduction, and acoustic radiation at the lips; side branches such as the piriform sinuses were not considered. Formant frequencies (F1 and F2) were determined by finding the peaks in the frequency response functions with an automated peak-picking algorithm (Titze et al., 1987). These operations were carried out separately for the coefficient grids corresponding to all three area function sets (A96, A08, and A9608x).

The resulting formant space for A9608x is shown in Fig. 3(b). With the exception of a small amount of overlap along the upper edge, this represents a one-to-one mapping between $[q_1, q_2]$ coefficient pairs and $[F1, F2]$ formant pairs, similar to those demonstrated in previous publications (e.g., Story and Titze, 1998; Story, 2005b). The formant pairs corresponding to the 11 vowel area functions in the A9608x set are shown as squares and circles. The formant grid outlines for the A96 and A08 sets are also shown in this figure as solid and dashed lines, respectively. Note that the A08 outline is shifted downward relative to that for A96, as would be expected based on the data shown previously in Fig. 1. More importantly, the formant grid based on A9608x encompasses nearly all of the space outlined by both the A96 and A08 sets.

4. Discussion

Finding the formant grid for the A9608x to be larger than that calculated for either the A96 or A08 sets suggests that the modes based on this combined set may be more useful for purposes of inverse mapping and vocal tract modeling. In the case of inverse mapping, a wider variation in $[F1, F2]$ formant frequencies can be accommodated within the formant space and thus related to a wider range of vocal tract area function shapes. For modeling vocal tract shape change as a component of synthesizing speech (cf. Story, 2005a), the modes from the combined set allow for production of a wider range of formant frequencies than those from the other two sets, and could potentially facilitate more natural sounding synthesis.

The similarity of the mode shapes across each of the three area function sets (A96, A08, and A9608x) does, however, raise the question of whether it would be possible to produce an expanded formant space for either the A96 or A08 sets simply by allowing the mode coefficient values to exceed their ranges produced by the PCA? For example, could the A96 modes and mean diameter function produce a formant space similar to the grid shown in Fig. 3 if the A9608x coefficient space was used instead of the more limited set of A96 coefficients? It is almost certain that expanding the coefficient ranges will have the effect of enlarging the formant space. In fact, Story (2004) used this approach to model a compensation for a lip-tube constraint imposed on the vocal tract shape. But expanding the coefficient range will always be limited by the degree of constriction produced within the vocal tract. That is, increasing the magnitude of either the $q_1$ or $q_2$ coefficients will ultimately result in occlusion of the vocal tract, and possibly generating unrealistically short or long tract lengths.

Acknowledgments

This research was supported by NIH Grant No. R01-DC04789.

References and links

1^F1 and F2 formant frequencies measured from audio recordings also indicated the same downward shift of F2 for the Story (2008) study relative to Story et al. (1996). The magnitude of the shift, however, was less than that observed in the calculated formants.

2^The use of 44 elements to represent an area function has been adopted by the author in previous publications, but is not a requirement. The number derives from the approximate spatial resolution obtained in MRI-based recon-
structions of vocal tract shape (e.g., Story et al., 1996). It is also convenient to use 44 elements for simulating speech with acoustic waveguide models because it allows for a sampling frequency of 44.1 kHz when the tract length is approximately 17.5 cm (typical adult male).

The area functions in the A96 set were indeed originally reported in Story et al. (1996) but were provided in a form with a variable number of sections (tubelets) where the length increment was constant across all vowels. Story and Titze (1998) resampled the area functions so that each area vector consisted of 44 sections, but the length increment was vowel dependent.

3Normalized distance is used for displaying the reconstructed area functions and modes because the actual vocal tract length is variable and depends on a given combination of mode scaling coefficients [see Eq. (7)].


