JASA LETTER TO THE EDITOR

Parametrization of the vocal tract area function using a subset selection approach (L)

Jorge C. Lucero[a]
Department of Computer Science, University of Brasília, Brasília, 70910-900, Brazil

ABSTRACT:
This letter introduces a parametrization of the vocal tract area function based on the position of a few points along the vocal tract. A QR decomposition algorithm is applied to area function data in various vowel configurations in order to identify those points with the most independent position patterns across vowels. Each point defines the shape of an associated kinematic region, and the overall area function is determined by the combination of the kinematic regions' shapes. The results show that only four data points, located at the tongue body, lips, and two at the tongue back, are enough to obtain accurate reconstructions of the vowels' area functions.
https://doi.org/10.1121/10.0005203

(Received 5 April 2021; revised 15 May 2021; accepted 20 May 2021; published online 9 June 2021)

[Editor: Anders Lofqvist] Pages: 4036–4038

I. INTRODUCTION

The area function of the vocal tract represents its cross-sectional area vs distance to the glottis. The function allows for the derivation of the acoustical and aerodynamic properties of the vocal tract and is an important component of physics-based synthesizers of vocal sounds [e.g., Story (2005a)]. In order to control and adjust the area function to different phonetic configurations and subjects, a formalization of its shape in terms of a few parameters is desirable and several techniques have been proposed [e.g., Atal *et al.* (1978), Mrayati *et al.* (1988), and Story and Titze (1998)].

Particularly, principal component analysis (PCA) has been widely applied to express the area function in terms of a basis of empirical eigenfunctions [e.g., Mokhtari *et al.* (2007), Story and Titze (1998), and Yehia *et al.* (1996)]. Each eigenfunction represents a principal mode of variation of the vocal tract shape, and the total area function may be described as neutral mean shape plus a linear combination of a reduced number of eigenfunctions. It has been claimed that the eigenfunctions characterize the degrees of freedom (DOFs) of the vocal tract and reflect the action of muscle synergies to achieve desired acoustic effects (Story, 2005b).

In this letter, an alternative characterization is proposed based on spatial DOFs. Suppose we want to infer the complete area function from information (values of cross-sectional area) of a few points along the vocal tract; then, which points should we choose?

The above question was addressed in the context of facial movement modeling during speech (Lucero and Munhall, 2008). In that work, position records of a set of markers on a subject's face were analyzed in order to identify those markers that follow the most independent motion patterns. The analysis was treated as a subset selection problem and was solved by a QR decomposition algorithm (Golub and Loan, 1996). Each independent marker defined a kinematic region of influence, and the total motion of the face was expressed as a linear combination of those regions' motions. Computer generated animations of the face were then produced by driving the independent markers with collected data.

The work of Lucero and Munhall (2008) work compared the subset selection approach with analyses of facial motion using PCA [e.g., Kuratate *et al.* (1998)] and argued that both approaches produce different but valid representations of the DOFs of the facial surface. However, while PCA extracts primary facial gestures, the spatial DOFs focus on the generating mechanisms of those gestures. Presumably, the kinematic regions are under control of individual muscles or synergies of muscles acting independently of each other, and therefore the spatial configuration of those regions should directly reflect the underlying biomechanical structure of the face.

Thus, the purpose of this letter is to introduce the subset selection approach for modeling the vocal tract area function, in the hope that the resultant parametrization might contribute to the understanding of shaping mechanisms and to the area function control in physics-based speech synthesis.

II. DATA AND PRE-PROCESSING

Data were taken from Story *et al.* (1996). These consisted of 10 sets of cross-sectional area values of the vocal tract at 0.396 cm regular intervals, from the glottal exit to the mouth. The area values were measured on MRI images of an adult male subject, and each set of images was taken while the subject was vocalizing a specific vowel: /i/, /ɪ/, /ɛ/, /æ/, /ʌ/, /ɑ/, /ɔ/, /o/, /ʊ/, and /u/.

Following Story and Titze (1998), all sets of area values were resampled to 44 data points by using cubic spline

[a]Electronic mail: lucero@unb.br, ORCID: 0000-0003-0597-3808.

interpolation. Thus, the area functions consisted of pairs (x_{ij}, y_{ij}), where x_{ij} is the distance from the glottis, y_{ij} is the area, $1 \leq i \leq 10$ is the vowel number, $1 \leq j \leq 44$ is the data point (point 1 is at the glottis and 44 is at the lips), and the x_{ij} values are equally spaced on the interval $[0, L_i]$, where L_i is the vocal tract length for vowel i (Fig. 1).

III. ANALYSIS

The application of the QR decomposition was described in detail in the work of Lucero and Munhall (2008). Briefly, assume a given $m \times n$ data matrix A. Then, A may be decomposed into factors in the form $AP = QR$, where P is an $n \times n$ column permutation matrix, Q is an $m \times m$ orthogonal matrix, and R is an $m \times n$ upper triangular matrix with non-negative diagonal elements (Golub and Loan, 1996). The first column of AP is the column of A that has the largest 2-norm, and the jth column of AP ($j > 1$) is the column of A with the largest component in a direction orthogonal to the directions of the first $j-1$ columns. As a result, the first k columns of AP (for some $k \leq n$) may be considered as a subset with the k most independent columns. The dimensionality of the data may be next reduced by fitting the remaining $n-k$ columns to the selected subset.

In the present case, the data matrix was constructed as follows. First, and following previous works (Mokhtari *et al.*, 2007; Story and Titze, 1998), all area values y_{ij} were converted into diameter values $z_{ij} = 2\sqrt{y_{ij}/\pi}$. As noted in the referenced works, the square root transformation has the effect of preventing negative area values resulting from the analysis and it also produces smaller errors when reconstructing the area functions from the selected subset. Further, working in the diameter domain has the additional advantage of keeping both coordinates of each data point in the same dimensional units.

Next, the mean across vowels was subtracted from each data pair, obtaining

$$(x^*_{ij}, z^*_{ij}) = (x_{ij}, z_{ij}) - (\overline{x_j}, \overline{z_j}), \tag{1}$$

where $\overline{x_j} = (1/10)\sum_{i=1}^{10} x_{ij}$ and $\overline{z_j} = (1/10)\sum_{i=1}^{10} z_{ij}$.

Finally, data were arranged as a matrix A with vowels in rows and data points in columns, in the form

$$A = \begin{bmatrix} X \\ Z \end{bmatrix}, \tag{2}$$

where submatrices $X = [x^*_{ij}]$ and $Z = [z^*_{ij}]$ have size 10×44.

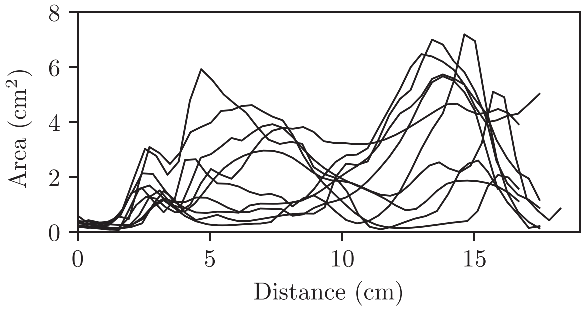

FIG. 1. Area functions.

The QR decomposition was applied to matrix A, and the first ten selected columns of A corresponded to data points 34, 44, 12, 19, 38, 7, 43, 30, 10, and 39, in that order (since there are only ten vocal tract data sets, then ten is the maximum possible number of independent columns).

After selecting a suitable number k of independent columns of A, the other columns were expressed as linear combinations of the selected ones. Then, distance-diameter pairs for each vowel i were reconstructed as

$$(x_{ij}, z_{ij}) = (\overline{x_j}, \overline{z_j}) + \sum_{\ell \in S} (x^*_{i\ell}, z^*_{i\ell}) c_{\ell j}, \tag{3}$$

where set S contains the indices of the selected k columns, and $c_{\ell j}$ are least squares fitting coefficients. Figure 2 shows examples of reconstructions of the area function from $k=3$ and $k=4$ points, obtained by converting the reconstructed diameter values back to area values. For $k=3$ there is a significant difference with the original data, but for $k=4$ the approximation is very close.

Coefficients $c_{\ell j}$ in Eq. (3) determine the effect of each independent point ℓ on the remaining points j of the vocal tract and define kinematic regions associated with each independent point. Figure 3 illustrates the kinematic regions in the case of $k=4$. Each plot shows the position of the selected independent point (black dot). For each plot, the control point was displaced from the mean position by 5 mm simultaneously in diameter and distance, and both in the

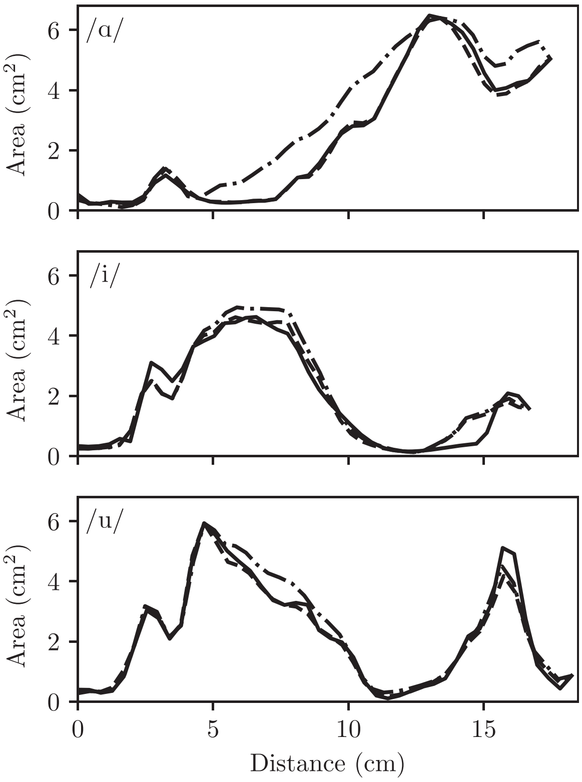

FIG. 2. Reconstructions of vocal tract area functions for vowels /ɑ/, /i/, and /u/, with $k=3$ (dash-point curves), $k=4$ (dashed curves), and original data (solid curves).

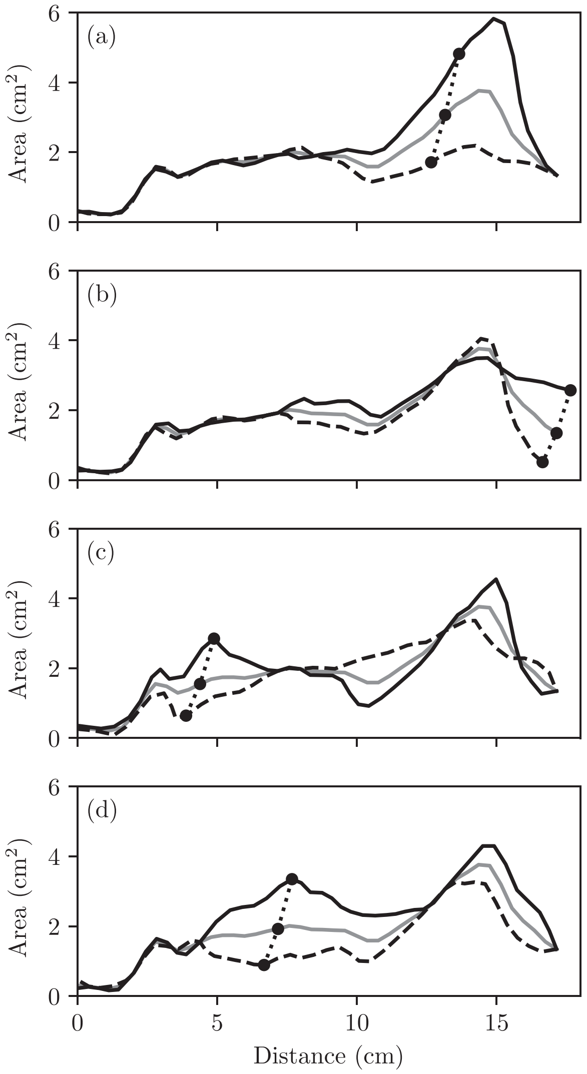

FIG. 3. First four independent points (black dots) and their effect on the area function shape, in order of importance from top to bottom. Gray curve: mean area function. Solid black and dashed curves: area functions that result when the control point is displaced to the indicated positions.

positive and negative directions, and the area function was recomputed in each case.

The first independent point [Fig. 3(a)] is located at the tongue body, and it controls the mouth cavity, without any effect on the mouth aperture. The second point [Fig. 3(b)] is at the lips, and it controls the mouth aperture and the area at the back of the mouth cavity. Also, the displacement in distance (horizontal axis) of this point changes the overall vocal tract length; whereas in case of the other points, their displacement causes a local distortion but not a total length variation. The third and fourth points [Figs. 3(c) and 3(d), respectively] are close to each other and located at the back of the tongue, and they both seem to relate to an arching-flattening motion of the tongue, and in the case of the third point, combined with an upward-downward motion. Together, these four points account for 96.4% of the variability of the diameter vs distance functions, computed following Story and Titze (1998), with mean rms errors of 1.8 mm for the cross section diameters and 0.6 mm for the section distances.

IV. CONCLUSION

The QR decomposition may be used to build a parametrization of the vocal tract area function in terms of the location of points which have the most independent patterns of variation across different vocal tract configurations. Only vowel configurations were considered, and the analysis showed that four points are enough to produce accurate reconstructions of the area functions.

As next steps, more comprehensive analyses using vocal tract data from various subjects as well as the inclusion of consonant configurations are planned to determine common features and individual variations of the resultant parametrizations. Also, the relation of each independent point and its associated kinematic region with the vocal tract acoustics should be investigated. Another possibility worth investigating is to drive the modeled area functions with time-varying articulograph data, in a fashion similar to those of the facial surface animations generated by Lucero and Munhall (2008).

ACKNOWLEDGMENTS

This research was supported by a grant from the Deans of Research and Innovation and of Graduate Studies of the University of Brasília (Brazil).

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (**1978**). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am. **63**(5), 1535–1555.

Golub, G. H., and Loan, C. F. V. (**1996**). *Matrix Computations*, 3rd ed. (The Johns Hopkins University Press, Baltimore), pp. 223–250.

Kuratate, T., Yehia, H., and Vatikiotis-Bateson, E. (**1998**). "Kinematics-based synthesis of realistic talking faces," in *International Conference on Auditory-Visual Speech Processing (AVSP'98)*, edited by D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson (Causal Productions, Terrigal-Sydney, Australia), pp. 185–190.

Lucero, J. C., and Munhall, K. G. (**2008**). "Analysis of facial motion patterns during speech using a matrix factorization algorithm," J. Acoust. Soc. Am. **124**(4), 2283–2290.

Mokhtari, P., Kitamura, T., Takemoto, H., and Honda, K. (**2007**). "Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients," J. Phon. **35**(1), 20–39.

Mrayati, M., Carre, R., and Guerin, B. (**1988**). "Distinctive regions and modes: A new theory of speech production," Speech Commun. **7**(3), 257–286.

Story, B. H. (**2005a**). "A parametric model of the vocal tract area function for vowel and consonant simulation," J. Acoust. Soc. Am. **117**(5), 3231–3254.

Story, B. H. (**2005b**). "Synergistic modes of vocal tract articulation for American English vowels," J. Acoust. Soc. Am. **118**(6), 3834–3859.

Story, B. H., and Titze, I. R. (**1998**). "Parameterization of vocal tract area functions by empirical orthogonal modes," J. Phon. **26**(3), 223–260.

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

Yehia, H. C., Takeda, K., and Itakura, F. (**1996**). "An acoustically oriented vocal-tract model," IEICE Trans. Inf. Syst. **79**(8), 1198–1208.