INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Functional principal component analysis of vocal tract area functions

Jorge C. Lucero

Dept. Computer Science, University of Brasília, Brasília DF 70910-900, Brazil
lucero@unb.br

Abstract

This paper shows the application of a functional version of principal component analysis to build a parametrization of vocal tract area functions for vowel production. Sets of measured area values for ten vowels are expressed as smooth functional data and next decomposed into a mean area function and a basis of orthogonal eigenfunctions. Interpretations of the first four eigenfunctions are provided in terms of tongue movements and vocal tract length variations. Also, an alternative set of eigenfunctions with closer association to specific regions of the vocal tract is obtained via a varimax rotation. The general intention of the paper is to show the benefits of a functional approach to analyze vocal tract shapes and motivate further applications.

Index Terms: vocal tract, area function, principal component analysis, functional data analysis

1. Introduction

A classical problem for physics-based simulation of voice and speech production is an adequate parametrization of the vocal tract shape [1, 2]. A model of the vocal tract shape and its associated acoustics is required in order to compute flow and pressure distribution inside it and obtain the resultant speech signal. For example, widely used models are wave reflection analogs [3] and transmission line circuits [4]. In both cases, the vocal tract is characterized as a sequence of concatenated elementary tubes, and their cross-sectional area is provided through an area function (i.e., cross-sectional area of the tubes vs. their distance to the glottis). Thus, a formalization of the area function in terms of a few parameters is desirable so that its shape may be easily controlled and adjusted to specific phonetic configurations and subjects.

Several parametrization strategies for the vocal tract shape have been followed in past studies. Articulatory models define the vocal tract geometry in terms of the position of individual articulators, such as the tongue, lips and velum [5]. An area function is then derived by appropriate conversion algorithms to cross-sectional area. Another strategy is to build a direct parametrization of the area function itself, and some techniques include: (1) concatenation of elementary curves defined in terms of physiologically relevant parameters, such as area and position of the maximum constriction, mouth opening, total vocal tract length [6]; (2) reduction of the vocal tract representation to a small number of tubes [7]; (3) expansion on a Fourier basis [8]; (4) expansion on an orthogonal basis of functions defined empirically by using principal component analysis (PCA) [1]. A review of various parametrization techniques may be found in Ref. [1].

This paper follows the study by Story and Titze [1], in which standard PCA was applied to a set of measures of various phonetic configurations [9]. By interpolating the data, each configuration was described as a vector with 44 components, where the components contained the cross-sectional areas of the vocal tract computed at regular length intervals from the glottis to the mouth. Variations in vocal tract length across the various configurations were initially disregarded, but they were added in a later study as an additional 45th component [10]. PCA allowed to decompose each area function as the sum of a mean function plus orthogonal components or modes. The mean function characterized a "neutral" vowel configuration with a formant structure similar to that of a uniform tube. The generation of the other vowels was then interpreted as perturbation of the neutral vowel, which was imposed through the orthogonal components.

Such a parametrization problem fits well into the framework of functional data analysis (FDA) [11]. FDA comprises a set of computational statistic tools for the analysis of patterns and variations in data expressed as sets of curves. In this context, a single datum is a smooth curve (i.e., an area function in the present case), and not a discrete numerical value. Some of those tools have already been applied in speech production studies; e.g., for analyzing articulatory movements [12] and aerodynamic patterns [13]. Here, a functional version of PCA (fPCA) will be applied to rework Story and Titze's parametrization [1] within the FDA context.

2. Data

The data consist of 10 sets of cross-sectional area values measured at 0.396 cm regular intervals, from the glottal exit to the mouth exit, from an adult male subject. Each set was computed from MRI images taken while the subject was vocalizing a specific vowel: /i/, /ɪ/, /ɛ/, /æ/, /ʌ/, /ɑ/, /ɔ/, /o/, /ʊ/ and /u/. Area values of the full data sets are available in Ref. [9].

3. Analysis

3.1. Functional form of the data

The first step of the analysis is to put the discrete set of measures into functional form. For each vowel i, with $i = 1, \ldots, 10$, the vocal tract area is given as a set of discrete pairs (x_{ij}, a_{ij}), for $j = 1, 2, \ldots, n_i$, where n_i is the number of samples, a_{ij} is the cross-sectional area at a distance x_{ij} from the glottis. Further, $x_{i0} = 0$ and $x_{in_i} = \ell_i$, where ℓ_i is the vocal tract length associated to vowel i. FDA assumes the existence of a smooth non-negative function $y_i(x)$ such that

$$a_{ij} = y_i(x_{ij}) + \epsilon_i, \tag{1}$$

where ϵ_i is an observational error or noise term.

Each area function y_i is defined over its own domain $[0, \ell_i]$ for the independent variable x. It is then convenient to define a common normalized variable s which ranges from 0 (glottal end) to 1 (mouth end), and consider the area functions as two-dimensional curves parametrized by s. Hence, we define the associated vector-valued area functions

$$\mathbf{y}_i(s) = (\ell_i s, y_i(\ell_i s))^T. \tag{2}$$

In addition, and in order to avoid negative area values arising from the analysis, a mapping $(0, \infty)$ into $(-\infty, \infty)$ is defined

http://dx.doi.org/10.21437/Interspeech.2017-181

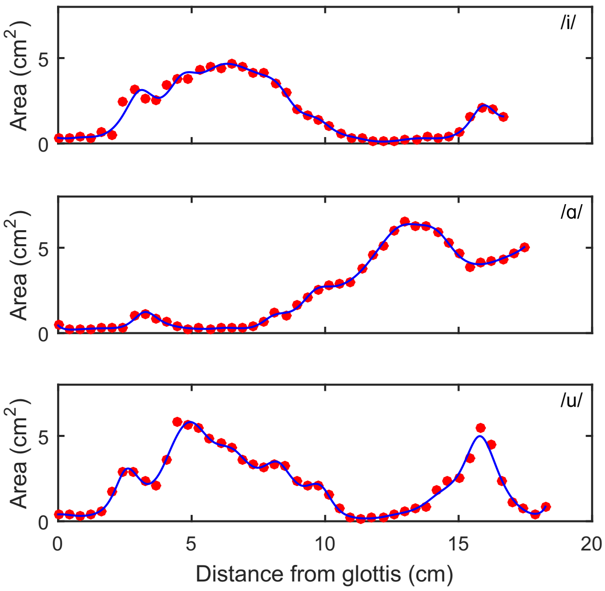

Figure 1: *Area functions. Red: measured data, blue: functional (smoothed) data.*

by $z_i(s) = \log(y_i(\ell_i s))$. The associated vector-valued log-area functions are

$$\mathbf{z}_i(s) = (\ell_i s, z_i(s))^T. \tag{3}$$

The log area functions are derived from the data by expressing functions z_i in a basis expansion form

$$z_i(s) = \sum_{k=1}^{m} c_{ik}\phi_k(s) \tag{4}$$

where $\phi_i(s), i = 1, 2, \ldots, m$ is a set of basis functions and c_{ik} are expansion coefficients. The expansion coefficients are computed by fitting the data to the above expansion with a standard least square method ($m < n_i$ is assumed).

Several types of basis may be applied to the above expansion; e.g., Fourier series, polynomial, regression splines and wavelet bases. Here, 2nd order B-splines (piecewise cubic polynomials) are adopted. This a commonly used general-purpose basis, numerically stable and with a continuous second derivative [14].

Fig. 1 shows examples of the resultant area functions for a basis of size $m = 25$. Fig. 2 shows the rms error of the fit for each vowel when varying the basis size. The error variations seems to stabilize at a basis size of 25, and therefore this size was adopted in the analysis. At this size, errors are below 0.05 cm^2 which is in the order of error values estimated by Story for his area measures [9]. Further, the resultant functions are visually smooth and approximate well the data.

3.2. Functional principal component analysis (fPCA)

As usual in PCA, the mean is first subtracted from the data. Each log area function is thus expressed as

$$\mathbf{z}_i(s) = \mathbf{z}_0(s) + \widetilde{\mathbf{z}}_i(s), \tag{5}$$

where $\mathbf{z}_0(s)$ is the mean

$$\mathbf{z}_0(s) = \frac{1}{10}\sum_{i=1}^{10}(\ell_i s, z_i(s))^T \tag{6}$$

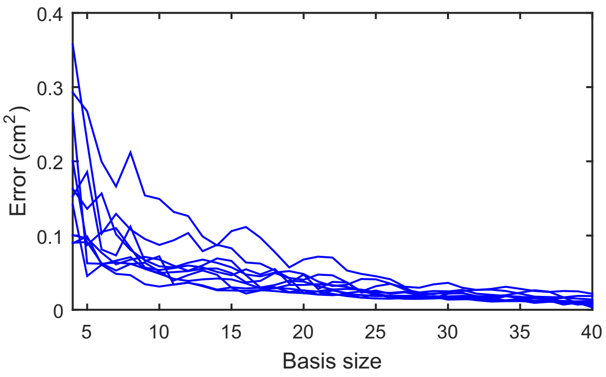

Figure 2: *Error of the basis expansion for each vowel.*

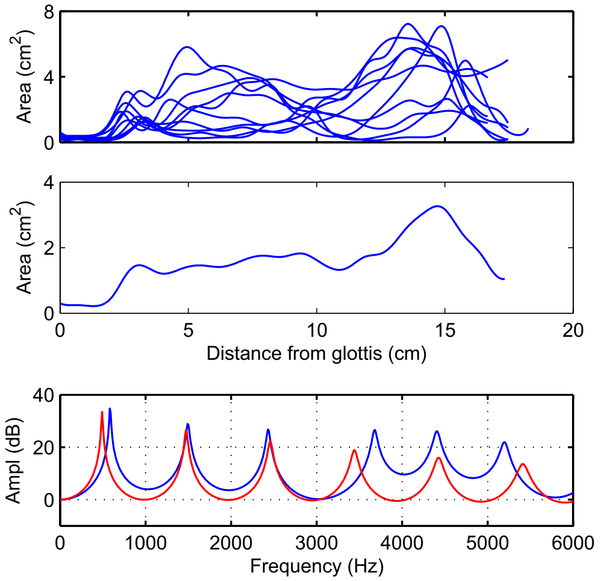

Figure 3: *Top: functional data set. Middle: mean area function. Bottom: frequency response of the mean area function (blue) and of a 1 cm^2 uniform tube of the same length (red).*

and $\widetilde{\mathbf{z}}_i(s)$ are the variations in relation to that mean.

Fig. 3 shows the functional data set and its mean area function. It also shows the frequency response associated to the mean area function computed by using a frequency-domain transmission line algorithm [15]. As noted by Story and Titze [1], this function produces a formant structure with almost equally spaced formants, similar to a uniform tube or a physiologically neutral vowel /ə/ configuration.

Functional PCA is defined as the following problem [11]:

1. Find eigenfunction $\mathbf{e}_1$ that maximizes the variance of the principal score $h_{i1} = \langle \mathbf{e}_1, \widetilde{\mathbf{z}}_i \rangle$ (inner product), with $i = 1, \ldots, N$ and N is the number of data functions (in the present case, $N = 10$), subject to the condition

$$\|\mathbf{e}_1\|^2 + \alpha\|D^2\mathbf{e}_1\|^2 = 1 \tag{7}$$

where $\|\mathbf{e}_1\|^2 = \langle \mathbf{e}_1, \mathbf{e}_1 \rangle$, D^2 denotes a second derivative, and α is a roughness penalty coefficient. The derivative term penalizes the size of the second derivative of $\mathbf{e}_1$, and has the purpose of avoiding excessive local variations in the eigenfunction. In the present analysis, a light smoothing of the results was obtained by selecting a penalty coefficient of $\alpha = 10^{-6}$.

2. Successively, compute eigenfunctions $\mathbf{e}_j$, $j = 2, 3, \ldots \ldots, k \leq N - 1$, that maximize the variance of $h_{ij} = \langle \mathbf{e}_j, \tilde{\mathbf{v}}_i \rangle$ subject to

$$\|\mathbf{e}_j\|^2 + \alpha \|D^2 \mathbf{e}_j\|^2 = 1, \tag{8}$$

and the orthogonality conditions

$$\langle \mathbf{e}_m, \mathbf{e}_j \rangle + \alpha \langle D^2 \mathbf{e}_m, D^2 \mathbf{e}_j \rangle = 0, \tag{9}$$

for $m = 1, \ldots, j - 1$.

In the bidimensional case, the inner product between two functions $\mathbf{e}_1 = (\xi_1, \chi_1)^T$ and $\mathbf{e}_2 = (\xi_2, \chi_2)^T$ over a domain $[0, 1]$ may be defined as

$$\langle \mathbf{e}_1, \mathbf{e}_2 \rangle = \int_0^1 \xi_1(s)\xi_2(s)ds + \int_0^1 \chi_1(s)\chi_2(s)ds. \tag{10}$$

Once the eigenfunctions and principal scores have been computed, the functional data is approximated in terms of the eigenfunctions as

$$\tilde{\mathbf{z}}_i(s) \approx \sum_{i=1}^{k} h_{ik} \mathbf{e}_k(s). \tag{11}$$

4. Results

All fPCA calculations were performed by using an FDA software package developed by Ramsay et al. [16]. Fig. 4 shows the first five eigenfunctions, which together account for 98.4% of the variance in the data, and Table 1 lists their associated principal scores for all vowels.

Table 1: *Principal scores for the first five eigenfunctions*

Vowel	Eig 1	Eig 2	Eig 3	Eig 4	Eig 5
/i/	−1.251	−0.204	−0.409	−0.097	0.004
/ɪ/	−0.471	−0.324	0.100	−0.008	0.096
/ɛ/	−0.255	−0.734	0.223	−0.167	−0.145
/æ/	0.098	−0.511	0.260	0.306	0.118
/ʌ/	0.537	−0.014	−0.051	0.112	0.052
/ɑ/	0.851	−0.134	−0.313	0.155	−0.120
/ɔ/	0.773	0.013	−0.156	−0.074	0.001
/o/	−0.011	0.558	0.250	−0.150	−0.198
/ʊ/	0.364	0.430	0.045	−0.382	0.195
/u/	−0.635	0.920	0.051	0.307	−0.004

The first eigenfunction has back-front symmetry and can be related to a backward-forward movement of the tongue. Precisely, its largest positive and negative scores occur for the back and front vowels /ɑ/ and /i/, respectively. The second eigenfunction describes an upward-downward movement or arched-flat tongue shaping combined with vocal tract length variation. In this case, the largest positive and negative associated scores appear for vowels /u/ and /ɛ/, respectively. Vowel /u/ demands a constriction in the mid-tract region and lengthening of the vocal tract (Fig. 1, bottom), whereas /ɛ/ demands an opposite shape variation.

The first two eigenfunctions replicate Story and Titze's results [1], except that length variations were missing in their analysis. Articulatory interpretations were not provided for the remaining eigenfunctions owing to their more complex shapes. In comparison, the present results are smoother which allows for a clearer analysis of the eigenfunctions' role. As Fig. 4 shows, the third eigenvalue mainly describes area expansion/contraction at the front region of the vocal tract and may reflect movements of the tongue tip. The fourth eigenfunction describes opening/closing of the mouth region and back region of the tongue, and may be related to movements of the jaw.

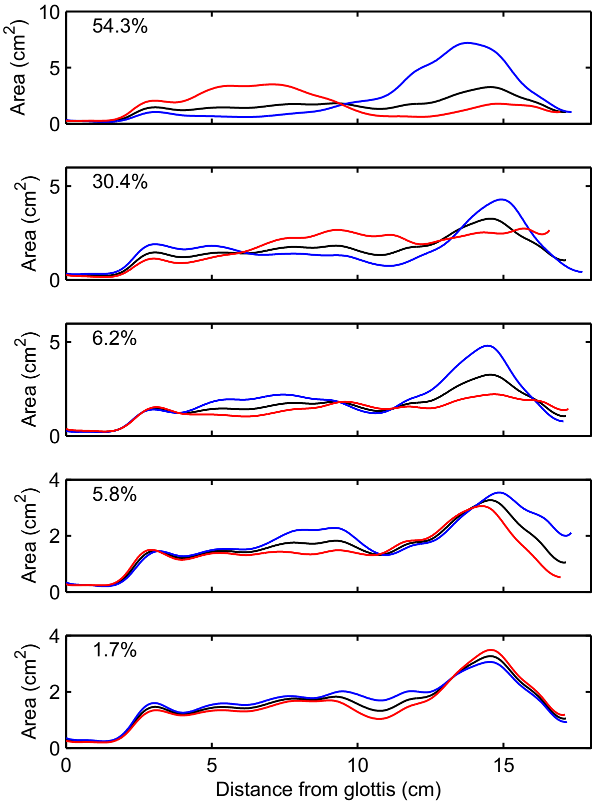

Figure 4: *Effect of adding and subtracting a multiple of each of the first five eigenfunctions (blue and red curves, respectively) to the mean area function (black curve). The percentage in each panel is the explained variance of the eigenfunction.*

The fifth eigenfunction describes more fine adjustments of the vocal tract shape and has a less significant role in terms of explained variance.

It is interesting to note that the tongue movements associated to the first four eigenfunctions match four of the parameters of Maedas's articulatory model [5], namely, backward-forward tongue position, arched-flat tongue shape, tongue tip position and jaw height. Let us recall that his model was built from factor analysis of vocal tract sagittal shapes. Jaw height was selected a priori as an articulatory parameter, and its influence was then subtracted from the data. Next, parameters related to the tongue shape were obtained by regular PCA. Vocal tract length variations were incorporated into additional parameters for lip protrusion and larynx height, whereas here they are embedded into the second eigenfunction.

An alternative set of eigenfunctions may be obtained by applying a varimax rotation (i.e., a rotation that maximizes the variability of the squared principal eigenfunctions). The re-

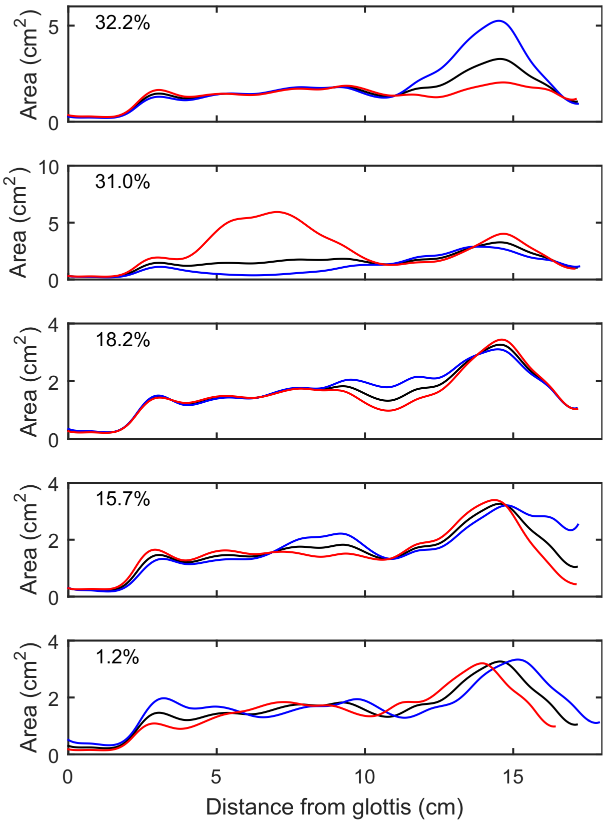

Figure 5: *Results of a varimax rotation of the first five eigenfunctions.*

sult are eigenfunctions that describe variability on more concentrated regions of the vocal tract, as shown in Fig. 5. The first four rotated eigenfunctions are directly related to variations in the front, back and mid vocal tract regions, and mouth opening respectively. The fifth rotated eigenfunctions captures variations of vocal tract length, combined with area variations at the epilarynx and pharyngeal regions. Note that the total variance explained by the set is the same as before, but that variance is more evenly distributed across the eigenfunctions. Nevertheless, the new set is still orthogonal. This alternative set may find application for controlling specific regions of the vocal tract, rather than modeling articulatory gestures.

Finally, the next two figures show results when reconstructing area functions by using the first five eigenfunctions. In Fig. 6, the area functions are approximated with rms errors of 0.21 cm^2 (/i/), 0.27 cm^2 (/ɑ/) and 0.40 cm^2 (/u/). Regarding the frequency responses, the reconstructions approximate well the formants except for the fourth formant of vowel /ɑ/. The increased value of this formant is caused by differences at the vocal tract exit in the reconstructed area function (Fig. 5, middle plot).

5. Conclusions

The main advantage of the functional approach to data analysis is that data is treated as continuous and smooth functions. Length variations of the vocal tract are easily incorporated into the analysis, and the vocal tract does not need to be sampled at regular intervals or even at the same positions across vowels. The results are also smooth functions and lend themselves to clear interpretations in physiological terms.

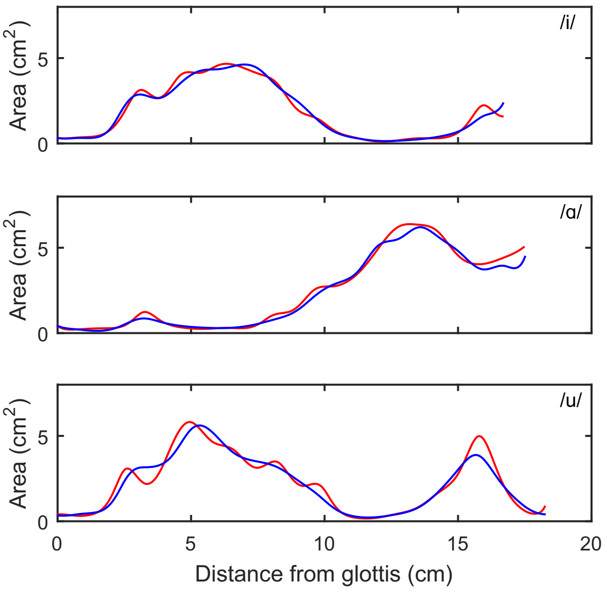

Figure 6: *Reconstructed area functions when using five eigenfunctions. Red: functional data from Fig. 1, blue: reconstruction.*

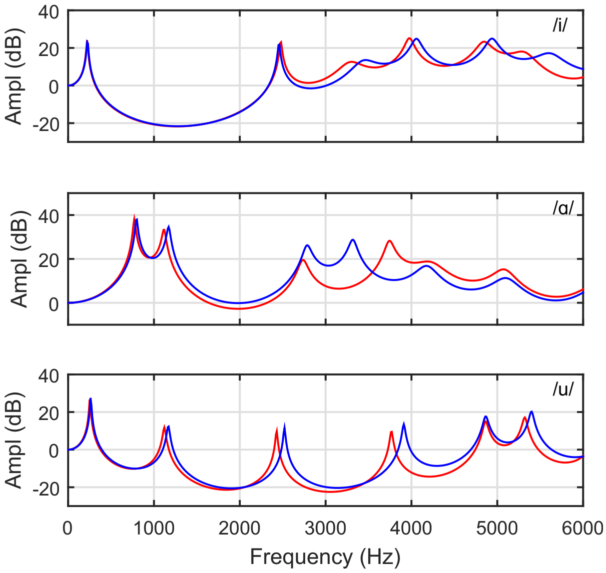

Figure 7: *Frequency response of the area functions in Fig. 6.*

The analysis has shown that vocal tract area functions for vowels are well characterized by a mean function plus five eigenfunctions. Area functions are reconstructed with low error and the first three formants are well approximated in all vowel configurations. Next research efforts will be dedicated to building functional models relating area function shapes to acoustic responses of the vocal tract.

6. Acknowledgments

This work was supported by CNPq/MCT (Brazil).

7. References

[1] B. H. Story and I. R. Titze, "Parameterization of vocal tract area functions by empirical orthogonal modes," *Journal of Phonetics*, vol. 26, pp. 223–260, 1998.

[2] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *Journal of the Acoustical Society of America*, vol. 117, pp. 3231–3254, 2005.

[3] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proceedings of the Fourth International Congress on Acoustics*, Copenhagen, 1962, pp. 1–4, paper G42.

[4] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal folds," *Bell Systems Technical Journal*, vol. 51, pp. 1233–1268, 1972.

[5] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*, W. J. Hardcastle and A. Marchal, Eds. Dordrecht: Kluwer Academic, 1990, pp. 131–149.

[6] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory to acoustic transformation in the vocal tract by a computer sorting technique," *Journal of the Acoustical Society of America*, vol. 63, pp. 1535–1555, 1978.

[7] M. Mrayati, R. Carré, and B. Guérin, "Distinctive regions and modes: a new theory of speech production," *Speech Communication*, vol. 7, pp. 257–286, 1988.

[8] J. Liljencrants, "Fourier series description of the tongue profile," *Speech Transmission Laboratory – Quarterly Progress Status Reports*, vol. 12, pp. 9–18, 1971.

[9] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *Journal of the Acoustical Society of America*, vol. 100, pp. 537–554, 1996.

[10] B. H. Story, "Vocal tract modes based on multiple area function sets from one speaker," *Journal of the Acoustical Society of America*, vol. 125, pp. EL141–EL147, 2009.

[11] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York: Springer-Verlag, 1997.

[12] J. O. Ramsay, K. G. Munhall, V. L. Gracco, and D. J. Ostry, "Functional data analyses of lip motion," *Journal of the Acoustical Society of America*, vol. 99, pp. 3718–3727, 1996.

[13] L. L. Koenig and J. C. Lucero, "Stop consonant voicing and intraoral pressure contours in women and children," *Journal of the Acoustical Society of America*, vol. 123, pp. 1077–1088, 2008.

[14] J. O. Ramsay and B. W. Silverman, *Applied Functional Data Analysis – Methods and Case Studies*. New York: Springer-Verlag, 2002.

[15] X. Zhou, Z. Zhang, and C. Espy-Wilson, "Vtar: A matlab-based computer program for vocal tract acoustic modeling," *The Journal of the Acoustical Society of America*, vol. 115, pp. 2543–2543, 2004.

[16] J. O. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and Matlab*. New York: Springer, 2009.