Vocal tract contour analysis of emotional speech by the functional data curve representation

Sungbok Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory,
University of Southern California
sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

Midsagittal vocal tract contours are analyzed using the functional data analysis (FDA) technique with which a vocal tract contour (VT) can be parameterized by a set of coefficients. Such a parametric representation of the dynamic vocal tract profiles provides a means for normalizing VT contours across speakers and offers interpretability of coefficient variability as the degree of contribution from specific vocal tract regions. It also enables us to examine the differences in VT behaviors as well as inter- and intra-speaker differences across different speech production styles including emotion expression. A set of FDA coefficients can be used as a feature vector of a given VT contour for further modeling. The efficacy of such feature vectors is tested using the Fisher linear discriminant analysis. A cross-validation accuracy of 65.0% was obtained in the task of discriminating four different emotions with combined data points from two speakers.

Index Terms: magnetic resonance imaging, vocal tract, emotion, functional data analysis, emotional speech production

1. Introduction

The acquisition and analysis of direct articulatory information of the *whole* vocal tract (VT) offer a means toward better insights into the underlying vocal tract shaping and its control. This allows the study of not only linguistically related articulation but paralinguistic articulatory movements such as those involved in emotional speech production. Recently we have developed a fast magnetic resonance (MR) imaging method [1] ("http://sail.usc.edu/span") which allows vocal tract imaging with a reconstruction rate of 21-frames per second with synchronized audio recording [2]. A region-based image segmentation algorithm for automatically marking the VT contours has also been developed [3]. This series of advancements in vocal tract data acquisition and image segmentation allows us to observe and track the entire upper and lower VT regions with a reasonable time resolution and, thus enables the study of vocal tract shaping dynamics simultaneously with the knowledge of corresponding speech acoustics.

Since it is now feasible to produce a large amount of MR images and VT contour data associated with various speech production styles or environments, it is important for us to develop automatic, or at least semi-automatic, vocal tract contour analysis methods in order to facilitate quantitative methods for speech production research. In this regard, we have explored the functional data analysis (FDA) technique[4] for processing VT contour data obtained during speech production under different emotion expression styles.

The FDA technique provides various statistical methods that are formulated exclusively to deal with curves (e.g., time series such as sensor trajectory data and vocal tract contours, etc.), not just individual data points. The FDA technique has been applied previously in several speech production research studies [5][6][7]. In this work, the technique is applied to process VT contours obtained from real time MRI [1] by representing them as functional data. This distinguishes the current study from other previous studies in which sensor trajectories of articulatory flesh-points have been dealt with. Specifically, this study focuses on investigating the differences in VT shaping associated with different emotion expressions such as anger, sadness and happiness with respect to normal, or neutral, speech emotion. By representing the continuous VT contours with a finite set of coefficients that are scalable, it is possible to represent them succinctly. The effectiveness of such a VT parametric representation using FDA is explored for capturing variation in articulatory behaviors associated with emotional speech production and some preliminary results are reported.

Section 2 describes our real time MRI experimental data of emotional speech, and the FDA approach used to parametrically model the VT contours obtained from that data. Section 3 describes our modeling results, Section 4 provides a discussion of the results and Section 5, our conclusions.

2. Method

2.1. Speech material

A set of four sentences, which are neutral in semantic content, were used for MR imaging with simultaneous speech audio recording. One female and one male native speaker of American English produced each sentence five times in a random order. Four different emotions, i.e., neutral, angry, sad and happy, were simulated by the subjects. The subject produced a set of 20 utterances for each emotion resulting in a total of 80 utterances (4 sentences x 5 repetitions x 4 emotions). The four sentences are: The doctor made the scar; foam antiseptic didn't help; Don't compare me to your father; That dress looks like it comes from Asia; and The doctor made the scar foam with antiseptic.

Simultaneous audio recordings during each MRI scan were collected using a fiber optical microphone. Each utterance was digitized in 16-bit amplitude resolution with a final 20-kHz sampling rate after a software cancelation of the typical MRI scanning noise made by the gradient coil [3].

In the current study each word "doctor" is acoustically segmented and the corresponding MR image sequence is analyzed. The word "doctor" was selected for the analysis because it has a stress in the first syllable and thus provides a common underlying speech production condition across different emotions. Therefore, the articulatory differences among different emotion expressions would be emphasized

predominantly in that word. A total 80 instances of the word "doctor" from the two subjects are analyzed in this paper.

2.2. MR image acquisition and tracking

The MR images were acquired using fast gradient echo pulse sequences and a 13-interleaf spiral acquisition technique with a conventional 1.5-Tesla scanner [1]. Those excitation pulses were fired every 6.856ms, resulting in a native frame rate of 11 frames per second (fps), that is, one entire frame of new information every 89ms (6.856ms x 13). Reconstruction of the raw data was implemented using a sliding-window technique with a window size of 48ms (the time that elapses between 7 successive excitation pulses). This produces a series of 68x68 pixel images with about 21 fps.

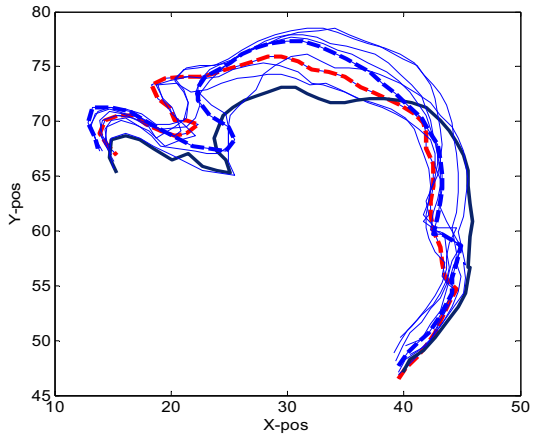

Figure 1. *An example plot of the lower VT contours of one production of "doctor" tracked from an MRI image sequence. Red and blue-dashed lines are the initial and final vocal tract contours, respectively. Black thick line is of the lowest tongue position (/ah/). The epiglottis has been removed in each VT contours.*

We have developed an MR image tracking method for unsupervised automatic region segmentation of an image using its spatial frequency domain representation as described in [3]. The algorithm was designed to process large sequences of real-time magnetic resonance (MR) images containing the 2-D midsagittal view of a human vocal tract airway. The segmentation algorithm uses an anatomically informed object model, whose fit to the observed image data is hierarchically optimized using a gradient descent procedure. One example of the vocal tract contours tracked from the sequence of MR images of a production of word "doctor" is shown in Figure 1 for the lower part of the vocal tract. For detailed mathematical description and the algorithm performance results, readers may refer to [3].

2.3. Functional data analysis of vocal tract contours

2.3.1. A brief review of the FDA curve representation

While most conventional statistical methods process a collection of individual data points, the FDA statistical framework is designed to process a collection of curves or functions [4]. The term "functional" reflects a view that by expressing discrete data in a functional form, one can better represent the underlying continuity of the physical or physiological system generating the data. Each curve is regarded as a sample of the underlying system. It also permits a more natural way to utilize its derivatives (e.g., velocity and acceleration) for system description or modeling.

In practice, such a functional representation of data is achieved by converting the raw sampled data points into a continuous functional form based on the expansion of basis functions with regularization with an appropriate order of derivative (usually a 3rd or 4th derivative),

$$F(x, y, \lambda) = \sum_j [x_j - y(t_j)]^2 + \lambda \int \left(\frac{d^3}{dt^3} y(t)\right)^2 dt$$

where x_j denotes observed value at time t_j in a discrete data sequence x, $y(t)$ is the function to be estimated from the observed sequence x, and λ is a smoothing parameter. And the function $y(t)$ is modeled as a linear combination of a set of basis functions,

$$y(t) = \sum_{k=1}^{K} c_k \varphi_k(t)$$

where $\varphi_k(t)$ is the k-th basis function with weight c_k and K is the number of basis functions. Therefore, by appropriately selecting λ and the order and number of basis functions, one can achieve a flexible approximation of discrete sampled data into a functional form. For detailed mathematical backgrounds and available software readers may refer to [4] and accompanying software which can be downloaded from the website "http://www.psych.mcgill.ca/misc/fda/index.html." The downloadable software was also used in this study.

2.3.2. Application to vocal tract contour representation

Since the FDA methodology requires one-dimensional representation of curves, each 2-dimensional VT contour is divided into x and y components and considered separately. This is a rather desirable condition for VT analysis because it provides an opportunity to investigate how each component is utilized or controlled by speaker for linguistic articulation as well as for speech emotion expression. In Figure 2, an example sequence of the x and y components of the VT contours of one production of word "doctor" (same token as shown in Figure 1) is plotted.

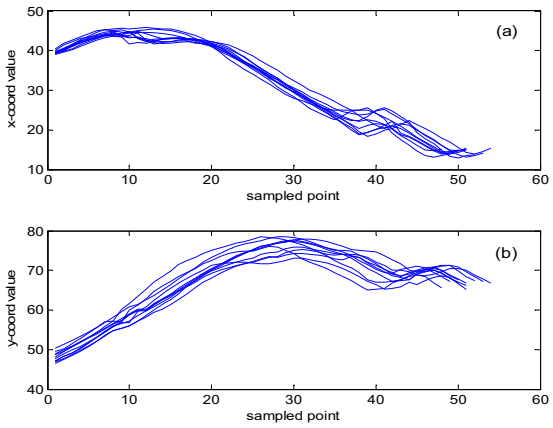

Figure 2. *An example of x and y component plots of one production of word "doctor" are shown in (a) and (b), respectively. The x-axis represents sampled points along the VT contours and the right-most position is of the lip.*

For a given x or y component curve, a set of 16 B-spline basis functions of order 5 with a λ value of 1E-12 is found to be reasonable enough to fit such a curve as shown in Figure 3. In the figure, an exemplary plot of the y-component and its FDA fit by the 16 coefficients are shown. Therefore, a set of

16 coefficients for a given component of VT contour can be regarded as a feature vector of that component.

Since the functional data representation of a curve utilizes a piecewise approximation of local segments of a curve delimited by break points, the coefficient sequence corresponds to the subsequent locations along the curve from the starting point. Therefore, by examining the variability of each coefficient as a function of emotion or as a function of subjects, one can specifically examine which region in the vocal tract is affected by emotion and by inter-speaker difference. This is one valuable merit of the FDA representation of VT contours. Another advantage is that by a common representation of VT contours with the same number of coefficients, VT contour normalization, irrespective of the VT length difference across speakers, can be achieved naturally.

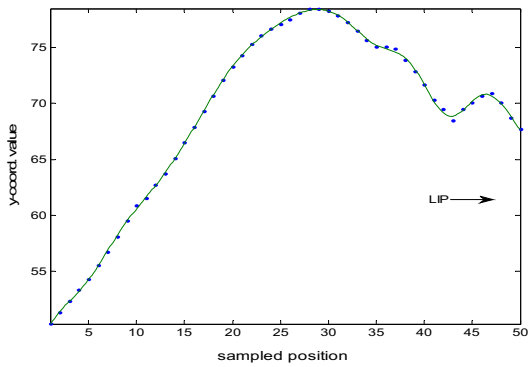

Figure 3. *Y-component values sampled along an example 2-D VT contour are fit by a corresponding set of 16 FDA coefficients. The fit captures local variations well.*

Only the lower VT contours (see Figure 1), covering the entire tongue contours, are analyzed in this study. The overall movements of the upper part of the vocal tract contour (representing the hard & soft palates, and the posterior pharyngeal wall) are relatively small and thus not considered in the current investigation. Finally, it is noted that, for the given number of basis and the order of B-spline functions, the choice of λ is critical to faithfully represent a given contour or curve. In fact, the choice of λ was found to be more important in data representation than the selection of the order and number of B-splines (cf. [7]).

2.3.3. Statistical analyses of vocal tract contours

Each set of 16 coefficients which corresponds to each *x* or *y* component of a VT contour is averaged across all the frames of each production of the spoken word analyzed and the mean VT contour and the standard deviation of each coefficient are computed. This procedure yields 80 representative VT components in each *x* and *y* dimension and the data of each dimension are subjected to one-way ANOVA analyses in order to test the significance of emotion effect to the variability of each coefficient as a function of emotion for each subject. More variability in coefficients indicates more susceptibility of the regions in the vocal tract represented by corresponding coefficients in emotional modulation.

In order to examine the effectiveness of emotion discrimination using the 16-coefficient representations of VT contours, Fisher linear discriminant analysis is performed for each *x* and *y* dimensions separately as well as for both dimensions combined. The discriminant analysis is also done for each speaker as well as for two speakers combined. It is noted that in the cases of combined subject analysis the z-scoring normalization has been applied to each subject's data before statistical analysis. All the statistical analyses are done with the SPSS statistical software package.

3. Results

The *x* and *y* components of averaged VT contours across the word sequences are shown in Figure 4 for subjects AB and EA as a function of emotion. Averaged variability of each coefficient is also shown in Figure 5 for both subjects. Because of space limit, only the case of *y* component of each subject is shown in both figures. Post-hoc tests show that at least one emotion differently behaviors when compared to the other emotions in both analyses of VT profile and coefficient variability.

3.1. Behavior of averaged vocal tract profiles

It can be observed from Figure 4 that the averaged articulatory component profiles are similar across subjects. Of course, this reflects the same linguistic environment (i.e., the same word "doctor" was produced by both subjects). Also the similar shapes as a function of emotion for both subjects imply that linguistic identity is preserved across different emotion expressions.

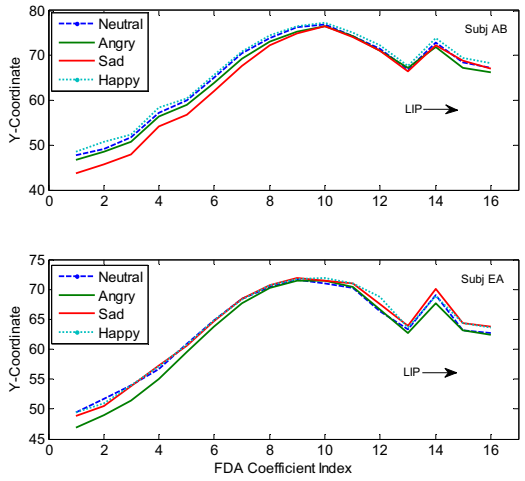

Figure 4. *Averaged x and y VT profiles expressed by 16 FDA coefficients are shown as a function of emotion. Speaker-dependent articulatory setting for emotion encoding can be observed.*

However, there exist some differences in averaged articulatory profile, or in *articulatory setting*, as a function of emotion. For subject AB, the sad profile shows a relatively lower tongue VT configuration when compared to other emotions. For subject EA, the angry emotion has the lower tongue VT configuration. This implies that somehow the lower pharyngeal region is affected by emotion coloring but there exists a speaker-dependent tendency.

3.2. Behavior in variability along the vocal tract

The variability of each coefficient reflects the activity of the corresponding segment in the vocal tract for linguistic realization as well as for emotion encoding.

It is observed from Figure 5 that the variability underlying normal linguistic articulations are severely affected by emotion coloring. Such a tendency is especially observable for

the back pharyngeal region for subject AB and for the front region of the vocal tract (i.e., the lip and tongue blade) for subject EA. This observation implies that there exist speaker-dependent strategies by which different parts of the vocal tract are utilized differently for different emotion encodings.

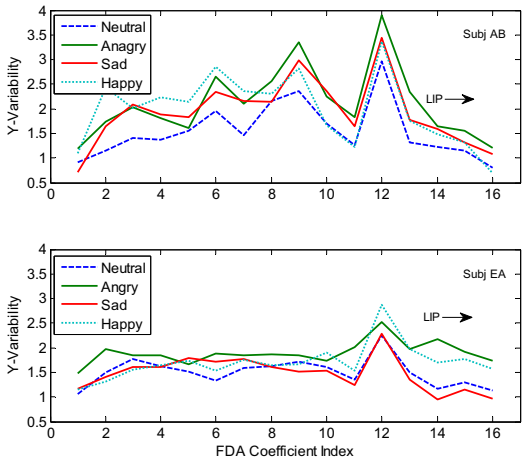

Figure 5. *FDA coefficient variability along the vocal tract is shown as a function of emotion. Speaker and emotion-dependent articulatory variability can be observed.*

3.3. Emotion discrimination power of parameterized vocal tract profiles

The leave-one-out, cross-validation accuracies of emotion discrimination using the FDA coefficients are shown in Table I. It is interesting to observe that VT contour component profiles of subject AB show much better emotion discrimination power when compared to those of subject EA, implying that subject AB utilizes more articulatory contrasts than subject EA in emotion encoding.

Although there exist some performance variability across speakers or VT components, the overall accuracy of 65% suggests that the FDA coefficient representation of VT profiles encapsulates a reasonable amount of emotion-specific articulatory information. In fact, this level of performance is comparable to results obtained with acoustic feature set [8]. Normalization effect of VT profile representation by the same number of coefficients across subjects may also contribute to this performance.

Table I. *Results of Fisher liner discriminant analysis: overall cross-validation accuracies (%) across four emotions.* The *16 FDA VT profile coefficients are used as independent variables.*

Subject	*Avg. x*	*Avg. y*	*Avg x+y*
AB	75.0	70.0	62.5
EA	60.0	50.0	47.5
Combined	46.3	56.3	**65.0**

4. Discussion

The main motivation of this study was to explore the FDA curve representation technique to develop a general-purpose tool for speech production research, especially for the processing of the large amount of vocal tract data obtained by the fast MR imaging acquisition and subsequent vocal tract contour tracking. The current results show that the FDA technique is promising toward achieving this goal. The normalization effect of the vocal tract size by the same number of parameters and the interpretability of FDA coefficients along the vocal tract seem to be good properties of the technique.

As by-product, this investigation also points to several detailed articulatory characteristics of inter and intra-speaker differences in emotion encoding in vocal tract shaping. The comparison of this articulatory domain behavior to the prosodic domain should provide a more comprehensive understanding of human speech emotion encoding. For instance, it is probable that inter-speaker differences in emotion encoding may arise from different weighting of the two independent domains. Also, the interplay between linguistic and affective aspects of speech articulation can be quantitatively studied. The scope of the investigation will be made broader with a larger number of speakers and by examining different speaking styles in the near future.

5. Conclusions

A relatively simple representation of VT contours by a finite number of coefficients using the FDA curve representation technique is found to be effective to describe the vocal tract behaviors. A parametric representation of the vocal tract profiles enables us to examine quantitatively the differences in VT behaviors as well as inter- and intra-speaker differences in different speech production styles such as different emotion expression. Such a set of coefficients can be used as a feature vector of a given VT contour for pattern comparisons. Efficacy of these feature vectors is tested using the Fisher linear discriminant analysis and a cross-validation accuracy of 65.0% was obtained in the task of four way emotion discrimination with the combined data points from two speakers. This is comparable to results obtained with large sets of pitch, energy and spectral features from speech acoustics. These results tell us that direct articulatory information carries crucial information about emotion modulation of speech. More investigations with different expressive styles of speech, from a larger number of speakers, is planned for the future.

6. Acknowledgements

This work was supported by grants from NSF and NIH.

7. References

[1] S. Narayanan, K. Nayak, S. Lee, A. Sethy, D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Amer., Vol. 115, 1771-1776, 2004.

[2] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," J. Acoust. Soc. Amer., 120(4):1791-1794, 2006.

[3] E. Bresch, and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," IEEE Transactions on Medical Imaging, Vol. 28, 323-338, 2009.

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

[5] J. Ramsay, K. Munhall, V. Gracco, D. Ostry, "Functional data analysis of lip motion," J. Acoust. Soc. Amer., Vol. 99, 3718-3727, 1996.

[6] J. Lucero, K. Munhall, V. Gracco, J. Ramsay, "On the registration of time and the patterning of speech movement," J. Speech Lang. Hear. Res., Vol. 40, 1111-1117, 1997.

[7] S. Lee, D. Byrd, and J. Krivokapic, "Functional data analysis of prosodic effects on articulatory," J. Acoust. Soc. Amer., 119(3): 1666-1671, 2006.

[8] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan. An acoustic study of emotions expressed in speech. In Proceedings of ICSLP, Jeju, Korea, October 2004.