An Exploratory Study of Emotional Speech Production using Functional Data Analysis Techniques

Sungbok Lee[1,2], Erik Bresch[1], Shrikanth Narayanan[1,2,3]

University of Southern California Viterbi School of Engineering,

[1]Departments of Electrical Engineering, [2]Linguistics, [3]Computer Science

sungbokl@usc.edu

Abstract. *Speech articulations associated with emotion expression were investigated using electromagnetic articulography (EMA) data and vocal tract data acquired using a fast magnetic resonance imaging (MRI) technique. The data are explored using functional data analysis (FDA) techniques for articulatory timing and vocal tract shape analyses. It is observed that the EMA trajectories of tongue tip movements associated with the production of a target word segment are quite similar across emotions examined in this study, suggesting that articulatory maneuvers for linguistic realization are largely maintained during emotion encoding. Results of the functional principal component analysis of the vocal tract shapes also support this observation. Mainly, the articulatory movement range and velocity (i.e., the manner of articulation) are modulated for emotional expression. However, another interesting articulatory behavior observed in this study is the horizontal shift of tongue tip positioning when emotion changes. Such a strategy by a speaker may facilitate the task of emotional contrast as long as such a variation of the place of articulation does not obscure linguistic contrast.* [Supported by NIH, ONR-MURI]

1. Introduction

In everyday conversation we comprehend and monitor not only *what* a talker says but also *how* the talker feels or what his or her attitude is. Our response to the other party is also conditioned not only by the literal meaning of the spoken words but also by the speaker's feeling or attitude behind it. It is reasonable to state that emotion expression and comprehension through spoken utterances is an integral part of speech communication.

Because of easy access to audio speech signal, the acoustic properties of emotional speech have been studied well in the literature (c.f., Scherer, 2003). Variations, or modulations, in pitch and amplitude patterns, as well as in segmental duration including pause, have long been known to be the major carriers of emotions. Acoustic correlates of some basic emotion categories (e.g., anger, sadness, and happiness) have also been well investigated in terms of pitch and energy as well as other temporal and spectral parameters such as segmental durations and spectral envelope features (Yildirim et al., 2004). Such knowledge could be useful for

developing speech applications such as machine synthesis and recognition (Lee and Narayanan, 2004). However, analysis of just acoustic features does not provide us a complete picture of the expressive speech production such as, for example, insights into the underlying vocal tract shaping, and their control, associated with emotion expression. Acquisition of direct articulatory information, although in general more cumbersome that speech recording, helps us tackle this issue to some extent. Recently we have collected emotional speech production data using an electromagnetic articulography (EMA) system as well as a fast magnetic resonance (MR) imaging technique. Notably, the MRI method allows vocal tract image acquisition with a rate of 22-frames per second with synchronized speech audio recording (Bresch et al., 2006; http://sail.usc.edu/span). This allows us to observe the entire midsagittal section of the vocal tract with a reasonable time resolution and thus study vocal tract shaping simultaneously with the corresponding speech signal. These speech production data are analyzed in the current study in order to explore the articulatory details of emotional speech, especially the question of how emotional articulation differs from a neutral articulation used for linguistic information encoding in speech.

To analyze the aforementioned multidimensional articulatory time series data, we utilize the functional data analysis (FDA) technique (Ramsay and Silverman, 2005). FDA provides various statistical methods that are formulated exclusively to deal with curves (e.g., time series such as EMA sensor trajectories, vocal tract contours, etc.), not just individual data points. The FDA technique has been applied in speech production research in several studies (Ramsay et al., 1996; Lucero and Koenig, 2000; Lee et al., 2006a). Specifically, we apply functional time alignment technique and functional principal component analysis to the articulatory data in order to investigate the differences in articulatory timing control and vocal tract shaping, respectively, between emotional and neutral speech articulations. Some preliminary results are presented in this report.

2. Acquisition of Speech Production Data

2.1. Speech material

A set of 4 sentences, generally neutral in semantic content, were used for both EMA data collection and MR vocal tract imaging. EMA data were collected from one male and two female subjects, and the male subject also took part in MRI vocal data collection. Subjects produced each sentence five times in a random order. Four different emotions—neutral, angry, sad and happy—were simulated by the subject. While such simulated emotion productions are known to be different from spontaneous unscripted productions, they are useful in providing a controlled approach to investigating some of the basic details (similar to the wide use of read speech in phonetic experiments). The 4 sentences are: (1) The doctor made the scar, foam antiseptic didn't help; (2) Don't compare me to your father; (3) That dress looks like it comes from Asia; (4) The doctor made the scar foam with antiseptic. In this paper, for each subject, the total 40 productions (2 sentences x 5 repetitions x 4 emotions) of the word "doctor" in sentences (1) and (4) were analyzed as a function of emotions.

2.2. EMA data recording

The Carstens AG200 EMA system was used to track the positions of three sensors in the midsagittal plane adhered to the tongue tip, the mandible (for jaw movement) and the lower lip. Reference sensors on the maxilla and bridge of the nose were tracked for head movement correction along with a sample of the occlusal plane of the subject acquired using a bite plate. The EMA system samples articulatory data at 200Hz and acoustic data at 16-kHz**.** Each sensor trajectory in the x-direction (anterior-posterior movement) and in the y-direction (vertical movement) with respect to the system coordinate is recorded by the EMA system. After data collection, each data trajectory was smoothed after correction for head movement and rotation to the occlusal plane.

2.3. MR vocal tract data acquisition

The MR images were acquired using fast gradient echo pulse sequences using a specially designed four-channel targeted phased-array receiver coil and a 13-interleaf spiral acquisition technique with a conventional 1.5-Tesla scanner (Narayanan et al, 2004). Excitation pulses were fired every 6.856ms, resulting in a frame rate of 11 frames per second (fps), and reconstruction of the raw data was implemented using a sliding-window technique with a window size of 48ms. This produces a series of 68x68 pixel images, each of which contains information from the preceding frame and a proportion of new information, thus affording us with an effective frame rate of 22 fps (i.e., one image every 46ms) for subsequent processing and analysis. We have also developed a method for synchronized, noise-mitigated speech recordings to accompany the MR imaging (Bresch et al, 2006).

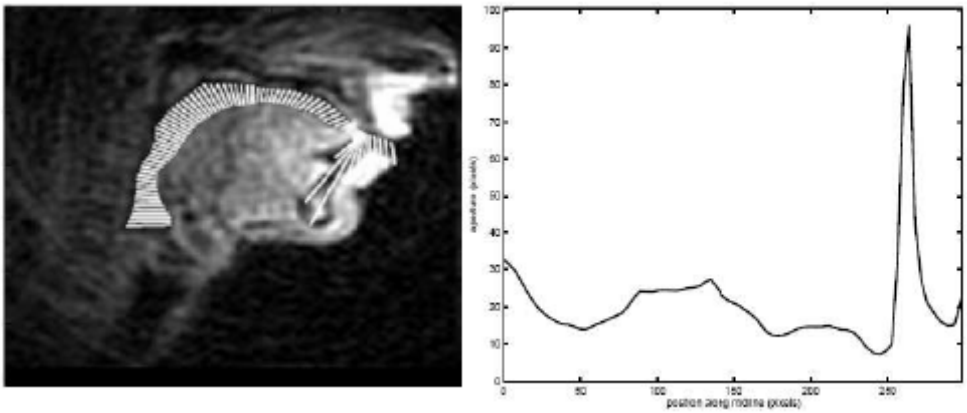

Figure 1. Examples of the vocal tract image and the outputs of the image tracking and the aperture function computation programs. The upper and the lower bounds of the vocal tract are determined by the MR image tracking program.

The raw MR images were tracked using a custom semi-automatic image tracking software based on the active contour model, or snake (Kass et al., 1987). The midsagittal contours of the speech articulators such as the tongue, the lips and the velum can be located by the program. We also developed an "aperture function" computation program which computes cross-sectional distances between the lower and upper boundaries of the vocal tract in the midsagittal plane from the larynx to the

lips (Bresch et al, submitted). An example MR image and the corresponding aperture function bounded by the upper and lower vocal tract contours are shown in Figure. 1. It is noted that the erroneously large cross-sectional distance around the jaw bone structure is an intentional artifact of the automated processing in order to capture the whole tongue contour by the image tracking system.

3. Data Analysis

Each production of the word "doctor" from sentences (1) and (4) (Sec 2.1) was segmented manually from the voice onset after /d/-closure to just before /m/ in the next word "made" by observing a speech waveform and spectrographic display. The beginning and end time stamps of the segment were used to bracket the corresponding tongue tip sensor trajectory in the EMA data. The same procedure was applied to speech waveforms that were simultaneously recorded during MRI sessions and the resulting time stamps were used to identify and collect image frames belonging to that segment. The tongue tip sensor trajectory data and tracked MR image frames were subjected to the functional data analysis techniques.

3.1. Functional data analysis

In order to apply the functional data techniques, the necessary first step is to convert sampled data points into a piece-wise continuous curve by a linear combination of a set of basis functions. This step also includes a smoothing, or regularization, procedure whose purpose is to reduce local random variations due to measurement errors. In the current study, two functional data analysis techniques are utilized: functional time alignment and functional principal component analysis. The former refers to an operation by which two signals of different lengths are being brought in phase with respect to one another. The alignment technique is used to examine the difference in articulatory timing across emotions with respect to the neutral speech. The latter is used to find dominant components in vocal tract shaping associated with the production of the target word segment. Matlab implementations of the FDA algorithms are publicly available at "ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns" and this study is based on that software.

3.2. Analysis of EMA tongue tip trajectories

After converting the raw tongue tip position trajectories into the functional data objects (i.e., curves) and computing the velocity patterns from them, a landmark based functional time alignment technique (Lee et al., 2005) was applied to the velocity curves as follows: First, a linear time normalization is applied to each individual velocity signal using the FDA smoothing and resampling methods. Then the control signals (i.e., velocity curves associated with neutral speech) are processed to get an optimized signal average as a reference signal. Each individual test signal is then time-aligned against the reference signal, and the corresponding time warping function is computed. Because a linear time normalization is done before FDA, purely linear time stretching is not captured; rather non-linear warping, which reflects local timing variations in tongue tip movements, is revealed.

3.3. Analysis of MR vocal tract contours

After tracking the contours of the tongue, the lips, the velum, and the pharyngeal wall up to near the laryngeal region, the vocal tract shapes delineated by the lower boundary of the vocal tract were subjected to functional PCA for each emotion. The functional PCA was performed for each coordinate separately and then the first few principal components in each coordinate were combined to restore the dominant modes of the vocal tract shape variations in the midsagittal section.

4. Results

4.1. Tongue tip movement

The analysis of the EMA data of emotional speech builds upon the preliminary descriptive results presented by Lee et al. (2005). In Figure 2, the EMA tongue tip position trajectories from the moment of acoustic release of /d/ to the offset of /r/ are shown for each subject. The most noticeable observation is that the shapes of trajectories are quite similar across emotions. This implies that the tongue tip movement for the linguistic realization of the target word is maintained by the speakers. Mainly, it is the tongue tip movement range and velocity (i.e., manners of the tongue tip movement) that are modified. Another mode of articulation that can be observed from subject AB and LS is a shift of the tongue tip positioning for /t/ when emotion changes. It is most clear for happy emotion of subject LS.

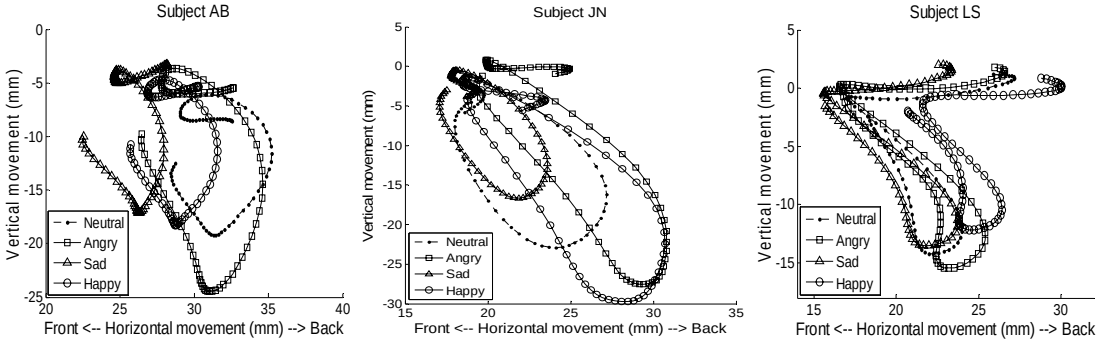

Figure 2. EMA tongue tip trajectories for each subject are shown from the moment of /d/ release (left end) to the offset of /r/ (right end). The left end is the start point. It is observed that trajectory shapes themselves are quite similar across emotions, suggesting the preservation of articulatory details related to linguistic contrast across different emotional expressions.

In Figure 3, the relative timing differences of velocity patterns of emotional speech with respect to the averaged neutral velocity signal are shown for subject AB as an example. It is observed that articulatory timing is quite stable for angry emotion but more variable for the other two emotions, especially for sad.

4.2. PCA analysis of the vocal tract shaping

A preliminary descriptive analysis of MRI data of emotional speech production by one subject was reported in Lee et al. (2006b). In this paper, we consider a quantitative functional data analysis of MRI data obtained from two subjects. In Fig.

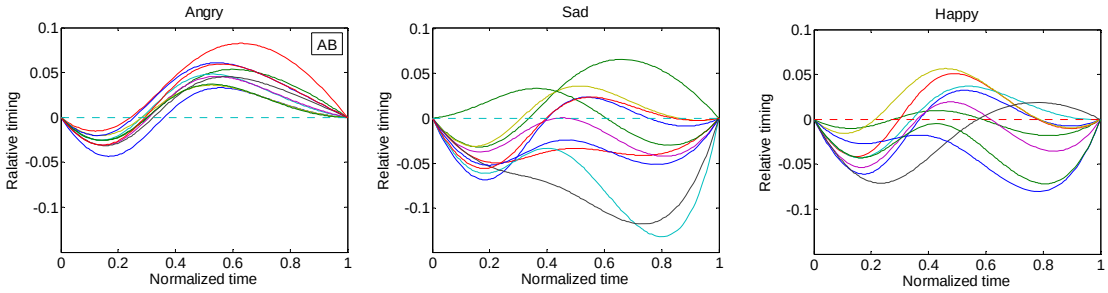

Figure 3. Relative timing of velocity patterns for each emotion category with respect to the neutral affect reference signal are shown after (linear) duration normalization for subject AB. Positive (negative) value means that an event occurs later (earlier) than in the reference signal.

4, as an exemplary illustration, the tongue shape variations associated with the first PCA component is shown for subject AB as a function of emotion type. On average across x and y coordinates, about 60% of variation in the data is explained by the first component. Similar mean tongue shapes and tongue shape variations across emotional categories confirms the finding from the EMA data that the tongue tip maneuvers associated with achieving the required linguistic contrasts is preserved in emotion expression. The second component was also found to show a similar tendency (although the plots are not shown here).

5. Discussion

The data from this study shows that articulatory maneuvers associated with achieving the underlying linguistic contrasts are largely maintained during emotional speech production. This is an expected result because (when) the primary purpose in speech is to render linguistic messages and the emotion-dependent articulatory modulation can be considered a secondary feature. More interestingly, however, this study has provided some insights to the question of *how* emotion-dependent articulatory modulations are realized during speech production. For the target word segment analyzed in the study, it was shown that the range and velocity of the tongue tip movements are the primary modulation parameters associated with emotional expression. Another possible modulation is a shift of the tongue tip positioning depending on the nature of emotion expressed by speakers. For instance, the data showed that subjects AB and LS modify the tongue tip positioning for /t/, especially for sad and happy emotions, respectively, when compared to their neutral speech. This can be interpreted as an emotion-dependent modification of anticipatory coarticulation from /t/ to /r/ in "doctor." It seems that the speakers have exploited the fact that the constriction location for /r/ in the oral cavity can be varied without affecting the phonetic quality of /r/ much. Based on those observations, it is reasonable to conjecture that speakers modulate the manner (e.g., articulatory

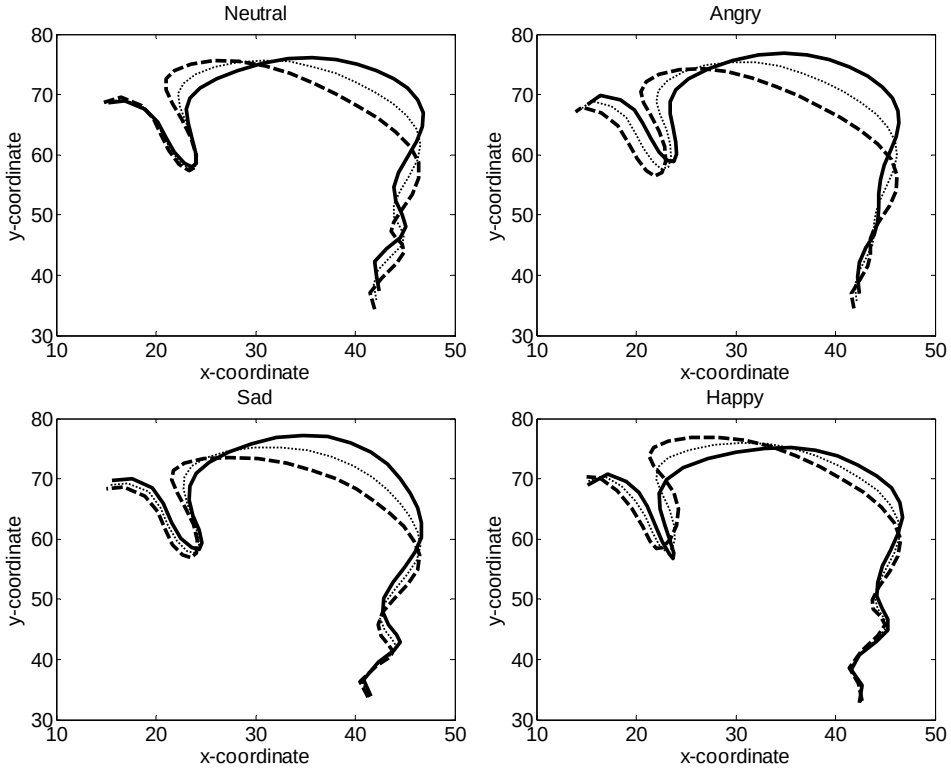

Figure 4. The tongue shape variation associated with the first PCA component are plotted as a function of emotion for subject AB. Dotted line represents each mean tongue shape. The tongue shape variations can be interpreted as dominant linguistic articulations modulated by emotional components.

movement range and velocity) and the place (e.g., tongue-tip positioning) of articulations as long as such modulations do not interfere with linguistic contrast.

Additionally, it was shown that articulatory timing control is also affected by the emotional encoding in a nonlinear fashion, although there exist no clear emotion-dependent or speaker-dependent patterns that can be discerned form our data. However, it appears that sad speech may exhibit more timing variability than other emotions, depending on speaker. Assuming a dynamical systems formalism, one could speculate that emotion-dependent control, either passive or active, of the stiffness associated with the tongue tip movement could be an underlying factor. It is noted that such data on timing control might be useful for an explicit modeling of articulatory timing for articulatory speech synthesis purpose. Finally regarding the utility of functional principal component analysis in conjunction with MRI derived time series of emotional speech, it was found that the first four components are sufficient to explain 95% of variations of the vocal tract shaping associated with the production of "doctor." In contrast, the result of conventional PCA indicates that the first seven components are needed to achieve the same level of performance. The functional PCA hence provides a more effective and simpler description of the vocal tract shaping.

The present study, while provides new insights into emotional speech production details, raises many questions notably between on the interplay

between the linguistic and affective aspects of speech production that need to be further investigated and validated. Such investigations are goals for our future work.

References

Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S. Synchronized and noise-robust audio recordings during realtime MRI scans. Accepted, *Journal of the Acoustical Society of America,* 2006.

Kass, M., Witkin, A., and Terzopoulos, D. Snakes: Active contour models. *International Journal of Computer Vision*, p. 321-331, 1987.

Lee. S, Bresch, E., Adams, J., Kazemzadeh, E., and Narayanan, S. A study of emotional speech articulation using a fast magnetic resonance imaging technique. International Conference on Spoken Language Processing, Pittsburgh, PA, 2006b.

Lee, S., Byrd, D., and Krivokapić, J. Functional data analysis of prosodic effects on articulatory. *The Journal of the Acoustical Society of America,* 119(3): 1666-1671, 2006a.

Lee, C. M., and Narayanan, S. Towards detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, 13(2):293-302, 2004.

Lee, S., Yildirim, S., Kazemzadeh, E., and Narayanan, S. An articulatory study of emotional speech production. In Proceedings of Eurospeech, Lisbon, Portugal, October 2005.

Lucero, J. and Koenig, L. Time normalization of voice signals using functional data analysis. *The Journal of the Acoustical Society of America,* 108, 1408-1420, 2000.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*, 115 (4), 1771-1776. 2004

Ramsay, J. O., Munhall, K. G., Gracco, V. L., and Ostry, D. J. Functional data analysis of lip motion. *The Journal of the Acoustical Society of America,* 99, 3718-3727, 1996.

Ramsay, J. O. and Silverman, B. W. Functional Data Analysis. 2nd Edition, Springer-Verlag, New York, 2005.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication 40(1-2), 227–256, 2003.

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. An acoustic study of emotions expressed in speech. International Conference on Spoken Language Processing, Jeju, Korea, 2005.