A Study of Emotional Speech Articulation using a Fast Magnetic Resonance Imaging Technique

Sungbok Lee, Erik Bresch, Jason Adams, Abe Kazemzadeh, Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering, University of Southern California, Los Angeles, California, USA
sungbokl@usc.edu

Abstract

A recently developed fast MR imaging system is utilized for a study of emotional speech production. Speech utterances and corresponding mid-sagittal vocal tract images are simultaneously acquired by the MRI system. Neutral, angry, sad and happy emotions are simulated by a male American English speaker. The MRI system and analysis results are described in this report. In general articulation is found to be more active in terms of the rate of vocal tract shaping and the ranges of spectral parameter values in emotional speech. It is confirmed that angry speech is characterized by wider and faster vocal tract shaping. Moreover, angry speech shows the more prominent usage of the pharyngeal region than any other emotions. It is also observed that the average vocal tract length above the false vocal folds varies as a function of emotion and that happy speech exhibit relatively shorter length than other emotions. It is likely that this is due to the elevation of the larynx and that may facilitate the higher pitch and larger pitch range manipulation to encode happy emotional quality by the speaker.

Index Terms: emotion, speech production, magnetic resonance imaging.

1. Introduction

Speech prosody (i.e., pitch and loudness modulation and durational modification) has long been known to be the carrier of emotion expressed in speech [1]. Acoustic correlates of specific emotional categories are also well investigated in terms of pitch, energy and temporal and spectral parameters [2]. However, the underlying articulations that govern the surface speech acoustics are not well studied when compared to their acoustic counterparts, largely due to the lack of non-invasive and fast vocal tract imaging method of the entire vocal tract during speech production. Although several articulatory studies exist which utilize point-tracking devices such as the electromagnetic articulography (EMA) [3][4], the scope is limited to a few points only along the parts of vocal tract that are accessible from outside. Accordingly, it can't provide information on the lower part of the vocal tract such as the pharyngeal configuration and the larynx height. Also, vocal tract imaging techniques based on ultrasound reflection at the boundary between tissue and air way can provide images of the tongue contour only [5].

Recently we have developed a fast magnetic resonance imaging method which allows vocal tract image acquisition with a rate of 21-frames per second with synchronized speech audio recording [6] ("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 to study vocal tract shaping simultaneously with the knowledge of corresponding speech acoustics. In this report, we apply the newly developed fast MR imaging method to study emotional speech production and present some preliminary results. To our knowledge, this is the first study of emotional speech production based on quasi-realtime MR images.

Some fundamental questions regarding emotion encoding in the articulatory domain include the following:

- how does a speaker modify speech articulation when emotion changes from neutral to other emotions;
- does there exist a common articulatory strategy across speakers to express a given emotion;
- if so, how can we characterize and model such common articulatory strategies across speakers for emotion coloring of speech sounds, both vowels and consonants.

In addition, it is also reasonable to hypothesize that speakers utilize both articulatory and prosodic aspects of speech to express their emotion effectively. This brings up additional interesting questions to be tackled including:

- is there any trading-off relation between the two modalities (e.g., vocal tract articulation vs. voice source activity) in the expression of emotions?
- is the weighting of two modalities for a target emotion expression emotion-dependent and/or speaker-dependent?

Exploring answers to these is important not only for theoretical reasons but also for practical purposes such as the development of an articulatory synthesizer that can handle appropriate emotional modulation of speech. This study is a preliminary effort toward answering the questions on the emotion encoding by human speakers by simultaneous observations of dynamic vocal tract shapes and the corresponding speech acoustics. Specifically, we investigate the dynamic changes of the midsagittal section of the vocal tract which is covered by the entire tongue contour.

2. Method

2.1. Speech material

A set of 4 sentences, which are mostly neutral in semantic content, were used for MR imaging with simultaneous speech audio recording. A 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 subject. 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 4 sentences are: (1) The doctor made the scar, foam antiseptic didn't help; (2) Don't compare me to your father;

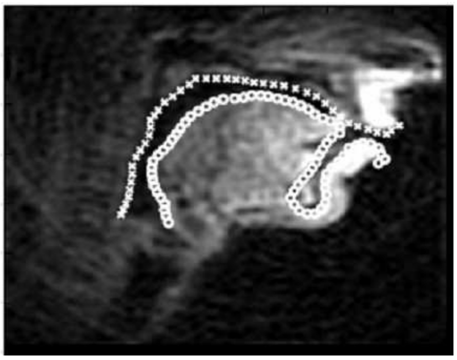

Figure 1: *An example of the vocal tract image and the corresponding output of the image tracking system. See "http://sail.usc.edu/span" for sample movies.*

(3) That dress looks like it comes from Asia; (4) The doctor made the scar foam with antiseptic.

Speech was recorded simultaneously with the MR image acquisition using a custom-designed audio-acquisition control box [8]. 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.

2.2. Vocal tract data acquisition

2.2.1. Vocal tract MR image acquisition

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 [6]. Those excitation pulses were fired every 6.856ms, resulting in a 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, each of which contains information from the preceding frame and a proportion of new information, thus affording us with an effective frame rate of 21 fps (i.e., one image every 48ms) for subsequent processing and analysis. A four-channel targeted phased-array receiver coil specifically designed for vocal tract imaging was used for this study. We have also developed a method for synchronized, noise-mitigated speech recordings to accompany the MR imaging [8].

2.2.2. MR image tracking

For image tracking, we have developed a semi-automatic image tracking software based on the active contour model, or snake [9]. After a manual determination of the contour of a target articulator at the first frame, the snake deforms itself to conform to the nearest salient "edge" by an iterative energy minimization process that balances between the complexity and the length of a contour based on intensity gradient across the probable articulatory contour.

The active contours model is modified to allow a contour to expand and contract from image to image seeking out an appropriate boundary. For this, we utilize the method of optical flow proposed in [10]. This method of optical flow uses second order differentials to approximate motion vectors. From the estimated motion vectors between frames we also can estimate the movement velocity of specified points along the boundary under consideration. An example MR image and the upper and lower vocal tract contours determined by the tracking system are shown in Fig. 1. It is noted that the deeply curved sub-lingual contour that circumvents the jaw bone structure is an intentional artifact in order to capture the whole tongue contour by the image tracking system. It can be removed by cutting out the section based on the fact that this lower portion of the contour does not vary much.

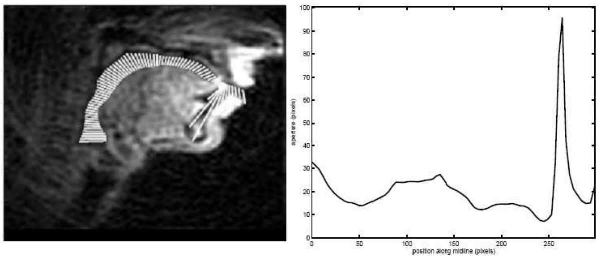

Figure 2: *Midsagittal cross-sections generated by the aperture function computation method and the resulting aperture function.*

2.2.3. Estimation of aperture function

Aperture function is defined by the series of cross-sectional distances between the lower and upper boundaries of the vocal tract in the midsagittal plane from the larynx to the lips. It depicts a vocal tract configuration and can be converted to the conventional area functions by an area conversion method for acoustic output simulation. It is noted that in the current study the unit of measure of the distance is *pixel*, not an actual length dimension.

As the first processing step, cubic splines are fitted to the two contours, and two tangents are found which connect both contours at the lip opening and the larynx. The midpoints of these tangents define the two endpoints of the midline of the vocal tract. Further support points of the midline are found using three repeated recursive bisections. As the next processing step, a temporary midline is obtained by fitting a cubic smoothing spline through the midline support points. Using the smooth temporary midline finely spaced cross sections of the vocal tract can now be found. This is done by intersecting perpendiculars to the midline with the two vocal tract contours. The final midline of the vocal tract can now be obtained by computing the midpoint of each cross section, and the aperture can be obtained as the length of the cross sections. Illustrative cross-sectional lines and the resulting aperture function approximately from the root of the epiglottis to the lips are shown in Fig. 2. The peak region near the lower incisor is due to the artifact of the image tracking system as mentioned previously.

3. Data analysis

Speech and vocal tract data for the word "doctor" in sentences (1) and (4) are examined in this report. The word has been investigated in a previous emotional speech production study using the electromagnetic articulography (EMA) system [7]. Five productions of the word as a function of emotion are analyzed in this report.

3.1. Acoustic analysis

First, the duration of the word "doctor" was measured manually from the voice onset after /d/-closure to just before /m/ in the next

Table 1: *Averaged duration and spectral parameter values for five "doc"[tor] productions as a function of emotion. Numbers in parentheses are range values.*

Emotion	dur (ms)	F0 (Hz)	F1 (Hz)	F2 (Hz)	F3 (Hz)
Neutral	352.7 (103.1)	151.6 (6.1)	789.9 (174.6)	1538.7 (259.8)	2645.5 (292.3)
Angry	454.3 (86.5)	249.9 (20.6)	881.8 (200.9)	1470.1 (178.4)	2634.5 (276.5)
Sad	448.3 (12.5)	190.6 (12.5)	735.3 (142.3)	1437.0 (354.6)	2517.2 (195.3)
Happy	396.4 (39.7)	264.7 (73.2)	832.0 (208.6)	1519.1 (378.5)	2694.9 (233.1)

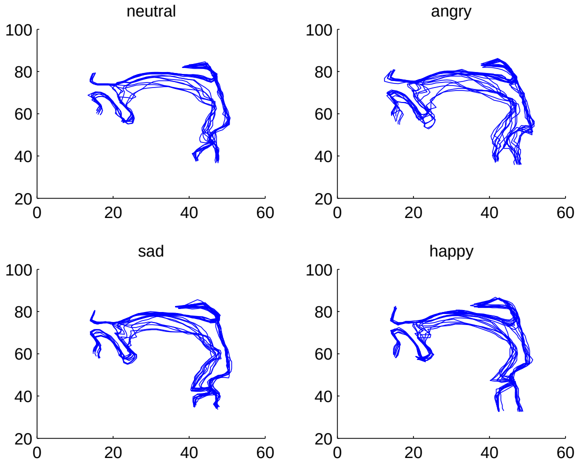

Figure 3: *Vocal tract outlines captured by the image tracking software are shown for an image sequence of "doctor" in each emotion.*

word "made" by observing waveform and spectrographic displays. Second, pitch and the first three formant frequencies were measured from the onset of /d/-release to the beginning of /k/-closure in "doc"(tor). Averaged pitch and the first three formants as well as maximum and minimum values were estimated in a given interval using the Praat software ("http://www.fon.hum.uva.nl/praat/").

It is noted that the current speech recordings were made in a noisy scan environment and it was necessary for us to remove the MR scanning noise whose spectrum spans acoustically important frequency bands for speech analysis through adaptive filtering schemes [8]. Through this pilot study, however, we confirmed that the noise reduction does not corrupt speech signal significantly and does not distort pitch and formant frequencies.

3.2. Vocal tract data analysis

We identified the MR image sequence corresponding to the word "doctor" in each utterance and visually examined each image sequence. We also measured the length of the vocal tract covered by the tongue in each frame of the image. The length is measured as a number of sections that covers from the minimum constriction location near the larynx (i.e., the position of the false vocal cords) to the point just before the beginning of the artificially induced peak in the aperture function. Also corresponding aperture functions are analyzed in terms of the frame-by-frame changes of the aperture functions, as a measure of articulatory activity, as a function of emotion. The rate computation is performed after a linear length normalization of each aperture function. Since the electromagnetic articulography (EMA) data for the tongue-tip movement are available for the current subject [7], the articulatory activity data are compared with the EMA data for that same word.

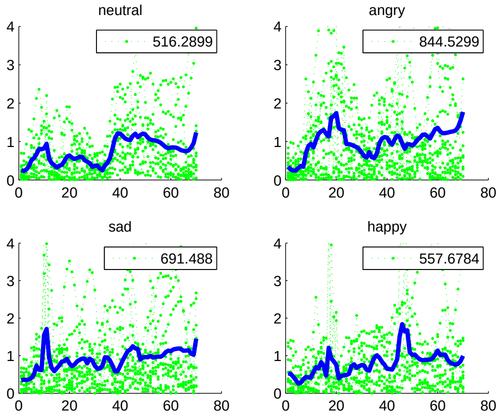

Figure 4: *Rate of change of the aperture function as a function of emotion. The x-axis denotes section number from the glottis and the y-axis represents the rate in "pixel." Thick lines represent averaged values across the whole image sequence. Numbers in the box represent averaged values across images and sections which is regarded as a measure of the activity of vocal tract shaping.*

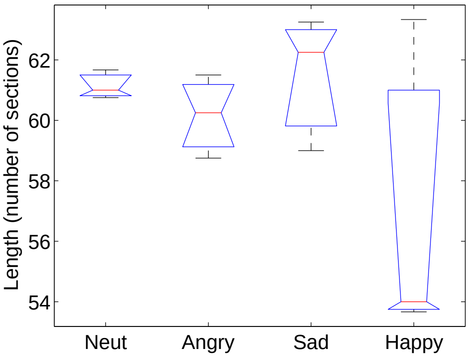

Figure 5: *Distribution of the vocal tract length covered by the tongue contour.*

4. Results

A summary of the duration, pitch and formant frequency measurements is shown in Table 1. Fig. 3 shows the outlines of the vocal tract are shown for an image sequence of the word "doctor" for each emotion. In Fig.4, the rate of change in midsagittal distances (i.e, aperture function) is plotted as a function of emotion. Fig. 5 shows the distributions of the vocal tract lengths are shown as a function of emotion.

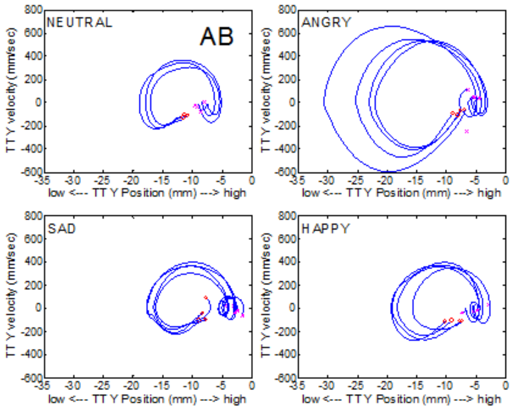

Figure 6: *Tongue-tip vertical movement by the current subject observed in a previous study [7] is shown for comparison purpose. Angry speech shows the most active tongue tip movement.*

Duration data indicates that segmental length for this speaker may increase when emotionally charged. The same tendency has also been observed from this speaker in a previous articulatory study using EMA point tracking [7]. It is observed that averaged pitch value and range are high for happy emotion which agrees with findings in the literature. It was also found that the range of F2 trajectory is particulary sensitive to emotional change.

It is visually clear from Fig.3 that the angry speech exhibits wider vocal tract shaping. It is also evident that the articulatory shaping in angry speech is most active not only in the oral cavity but also in the pharyngeal region. In fact the current speaker is utilizing the pharyngeal region most prominently in angry speech. This visual observation also can be confirmed in Fig.4 in which the rate of change of the aperture function is plotted as a function of emotion. In the case of happy speech, as can be observed in Fig.3, the position of the false vocal folds (i.e., the maximum constriction position near the larynx) is higher than for any other emotion, and thus the overall active vocal tract length covered by the tongue contour is shorter in happy speech as can be observed in Fig.5. It is speculated that this particular vocal tract configuration of the happy speech may contribute to the large pitch and F2 ranges and higher F3 for the expression of that emotion as can be observed in Table 1.

In Fig. 6, the tongue-tip movement data obtained by the electromagnetic articulography (Carstens Ag200) system are shown for comparison purpose as the data have been obtained from the same speaker. Clearly, angry speech shows the most active tongue tip movements in terms of movement range and velocity. This tendency may agree with the observation from Fig. 4.

5. Discussion

By examining the whole vocal tract images associated with emotion speech production we were able to confirm the previous findings that were based on limited access of the vocal tract with the EMA system and to extend the scope of emotional speech production study into the back part of the vocal tract. We found that the articulation of angry speech is active not only in the oral cavity but also most prominently in the pharyngeal region. This may be related to the observation that angry speech exhibits the largest speech rms energy for this speaker, which requires a larger air volume throughout. Also there seems evidence that the larynx elevation is one mechanism that could be involved in the articulatory realization of happy emotion. Without the availability of the recently developed fast MRI method described in this study, such observations wouldn't be possible.

We are aware of the fact that the degree and manner of emotional articulation for a target emotion are dependent on speaker and/or emotion quality realized. As we collect more emotional speech production data from more subjects, the current findings in articulatory encoding of emotion will be verified against other speakers and it is hoped that we can observe some common characteristics of emotional articulation across speakers. The goal for our future work is to extend this initial study to a larger set of emotional speech production data as well as to model the underlying articulatory-acoustic relations.

6. Acknowledgements

This work was supported by grant from NIH. We also thank for the help from USC-SPAN team.

7. References

[1] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," Speech Comm., 40, 2003.

[2] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," IEEE Trans. on Speech & Audio Processing, 13(2), 293-303, 2005.

[3] D. Erickson, O. Fujimura, B. Pardo, "Articulatory correlates of prosodic control: Emotion and emphasis," Language and Speech, 41, 395-413, 1998.

[4] M. Nordstarnd, G. Svanfeldt, B. Granstrom, D. House, "Measurements of articulatory variation in expressive speech for a set of Swedish vowels," Speech Communication, 44, 187-196, 2004.

[5] Y. Akgul, C. Kambhamettu, M. Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences," IEEE Comp. Vision and Pattern Recog., 124:298-303, 1998.

[6] 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., 115:1771-1776, 2004.

[7] S. Lee, S. Yildirim, S. Narayanan, Abe Kazemzadeh, "An articulatory study of emotional speech production," EUROSPEECH, Lisbon, Portugal, 2005.

[8] E. Bresch, J. Nielsen, K. Nayak, S. Narayanan, "Synchronized Audio Recording during Real-Time MRI Scans," J. Acoust. Soc. Am., 2006 (submitted)

[9] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," Int. J. Comput. Vis., vol. 1:321-331, 1988.

[10] B. Lucas, T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," Seventh Int. Joint Conf. Artif. Intell., Vancouver, Canada, 674-679, 1981.