USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging

Jangwon Kim, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Sungbok Lee, and Shrikanth Narayanan

University of Southern California, Los Angeles, CA, U.S.A.
jangwon@usc.edu, toutios@sipi.usc.edu, yoonckim@usc.edu,
yinghuaz@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu
http://sail.usc.edu/span, http://mrel.usc.edu

Abstract

This paper introduces a new multimodal database of emotional speech production recorded using real-time magnetic resonance imaging. This corpus contains magnetic resonance (MR) videos of five male and five female speakers and the results of evaluation for the emotion quality of each sentence-level utterance, performed by at least 10 listeners. Both speakers and listeners are professional actors/actresses. The MR videos contain MR image sequences of the entire upper airway in the mid-sagittal plane and synchronized speech audios after noise cancellation. The stimuli comprises the "Grandfather" passage and seven sentences. A single repetition of the passage and five repetitions of the sentences were recorded five times, each time with a different acted emotion. The four target emotions are anger, happiness, sadness and neutrality (no emotion). Additionally one repetition of the Grandfather passage was recorded in a neutral emotion and fast speaking rate, as opposed to a natural speaking rate for the rest of the recordings. This paper also includes a preliminary analysis of the MR images to illustrate how vocal tract configurations, measured in terms of distances between inner and outer vocal-tract walls along the tract, vary as a function of emotion.

Keywords: emotional speech database, magnetic resonance imaging, emotional speech production, emotion evaluation.

1. Introduction

Previous studies in emotional speech production have reported evidence that different emotions affect articulatory movements differently in a systematic way. Erickson et al. [1, 2] reported that the positions of the tongue tip, the tongue dorsum, the jaw and the lips, can be characterized by emotion type of speakers. Lee et al. [1] also found that emotional speech articulation exhibits more peripheral or advanced tongue positions, especially for sadness, which is in the line with the finding of Erickson et al. Lee et al. [3] also found that the movement range of the jaw is largest for anger, which is also in the line with the finding of Erickson et al. Kim et al. [4, 5] reported that the position, movement range and speed of articulators are important cues for distinguishing emotional state of speakers of their dataset. Kim et al. [6] also found that angry speech involves more emphasis on articulatory variation than happiness, while happiness involves more emphasis on the pitch variation; the degree of relative emphasis varied depending on speakers. Although the findings of these preliminary studies are limited to small lexical contents and small number of subjects, they hint that emotional information is encoded in a systematic variation of articulatory movements and prosodic behaviors.

The articulatory information used in the aforementioned preliminary studies was obtained using EletroMagnetic Articulolgraphy (EMA). EMA allows monitoring the movements of a few flesh-point sensors attached on articulatory points of interest. Although EMA provides information of articulatory motions at a relatively fast sampling rate (100, 200 and 400 Hz for the NDI Wave system) with reliable accuracy overall [7] the number of articulatory points (maximum 6 sensors for a single NDI Wave system), and the vocal-tract region that the EMA can monitor are limited.

Recently, real-time magnetic resonance imaging (rtMRI) [8] with simultaneous speech recording [9] were utilized to study emotional variation in the articulatory domain. The rtMRI offers the entire view of the upper airway of a speaker in any scan plane of interest with no need of repetition, and therefore dynamic information on the movements of the lips, the jaw, the velum, the tongue (including in the pharyngeal region), and the larynx in the mid-sagittal plane can be captured. This information provided by the rtMRI allows us to study their patterns of articulatory coordination and global vocal tract parameters, such as the vocal tract length. Analyzing the rtMRI data of emotional speech of a male speaker, Lee et al. [10] found that angry speech is characterized not only by wider and faster vocal tract shaping, but also by more usage of the pharyngeal region than other emotions (neutrality, happiness and sadness). They also reported that happy speech exhibited shorter vocal tract length than the other emotions in the paper.

One of the important characteristics of emotional speech production is inter- and intra-speaker variability. Such a largely heterogeneous nature of emotion expression and perception has been raised in literature (e.g., [11, 6, 12, 13]), but the knowledge in orchestrated articulatory control for emotion encoding on top of linguistic articulation is still limited. Another interesting aspect is the mismatch between perceived emotion of listeners and the emotional state of speakers.

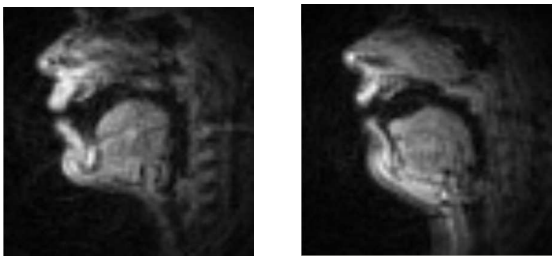
The present paper introduces a new speech production corpus of emotional speech, namely the USC-EMO-MRI corpus, which was collected at the University of Southern California using rtMRI. This corpus comprises the entire mid-sagittal vocal tract images and simultaneously recorded speech audios of 10 speakers, and emotion quality labels by at least 10 listeners for each speaker's data. The articulatory and acoustic data are collected with three basic emotions (happiness, sadness anger) and neutrality (no emotion presented).

The USC-EMO-MRI corpus is designed as a resource for systematic analysis for the inter- and intra-speaker variability

of emotional speech in the articulatory movements and prosodic behaviors. This corpus also aims at assisting more comprehensive modeling of the vocal tract shaping in the upper airway and eventually the *joint* modeling of articulatory and acoustic behaviors with human-like emotion coloring. In addition, this corpus could be equally useful for other speech production studies, such as articulatory-to-acoustic forward mapping [14] or acoustic-to-articulatory inversion [15, 16] of emotional speech.

2. Data collection

Sequences of mid-sagittal images of the upper airways of five male and five female subjects were recorded in a 1.5 Tesla MRI scanner using a custom upper airway receiver coil and a real-time MRI acquisition protocol that has been described in detail elsewhere [17]. The resolution of the MR image is 68 × 68 pixels, where the pixel size is 3 mm × 3 mm. Hence, the field of view is 204 mm × 204 mm. All subjects are professional actors or actresses who have had theatrical vocal training. Figure 1 shows MR images extracted from the rtMRI videos for one male and one female speaker in the database.



(a) an MR image of M1 (b) an MR image of F1

Figure 1: *Example MR images of two subjects, M1 and F1 in Table 1*

The frame rate of the reconstructed MRI video is 23.180 frames/sec. The details of MR image reconstruction that we followed for this database have been described in [18]. Speech audio was simultaneously recorded, using a custom fiber-optic microphone at a sampling rate of 20 kHz during the MR imaging. Noise cancellation was performed on speech audio using a custom adaptive signal processing algorithm [9], the audio and video signals were synchronized. See [9] for the details of noise cancellation and audio-video synchronization that we followed.

The subjects were asked to read stimuli after immersing themselves in one of four target emotions, such as neutrality, happiness, anger and sadness, in the scanner. The stimuli consist of the "Grandfather" passage and seven sentences.

- The "Grandfather" passage: You wished to know all about my grandfather. Well, he is nearly ninety-three years old; he dresses himself in an ancient black frock coat, usually minus several buttons; yet he still thinks as swiftly as ever. A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a trifle. Twice each day he plays skillfully and with zest upon our small organ. Except in the winter when the ooze or snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language.
- 7 sentences:
  1. John bought five black cats at the store.
  2. The Leopard, skunk and peacock are wild animals.
  3. Charlie, did you think to measure the tree?
  4. The queen said the KNIGHT is a MONSTER.
  5. Hickory dickory dock, the mouse ran up the clock. Hickory dickory dock.
  6. 9 1 5 (short pause) 2 6 9 (short pause) 5 1 6 2.
  7. Ma Ma MA (short pause) Ma Ma Ma (short pause) Ma Ma Ma Ma.

The subjects uttered the passage in normal speaking rate for each emotion and a second time in fast rate only for neutrality. The subjects also uttered six or seven sentences with seven repetitions in only normal speaking style for each emotion. The order of the sentences was randomized at each repetition of data collection, except for the sentences 6 and 7. The subjects were asked to repeat their intonation of the sentence 6 when they read the sentence 7. The sentence 7 was always presented right after the sentence 6. Subjects were asked to emphasize "KNIGHT" and "MONSTER" when reading the sentence 4. Table 1 shows the sentences uttered by each subject. The set of sentences is designed to investigate the effects of emotion expression on the syntactic, prosodic and rhythmic structure in corresponding spoken utterance, including one reiterant speech.

3. Emotion evaluation

The emotion quality of the data was evaluated for each sentence-level utterance with perceptual tests from at least 10 actors/actresses. After listening to speech audio, the evaluators were asked to give their opinion on three questions: (1) the best representative emotion among five categories, such as neutrality, anger, happiness, sadness, and 'other,' where 'other' was for the case that none of the listed four emotions was the best, (2) confidence in their evaluation, and (3) the strength of emotion expression. Confidence and strength were evaluated on a five-point Likert scale. The best emotion was determined by majority voting. If there were multiple emotions with the same evaluation score, the one of higher mean of confidence scores was chosen.

Table 1 shows the number of evaluators, sentences, the average and standard deviation of matching ratio between target and evaluated emotions. 'Sent' 1-7 denotes all seven sentences were recorded, while 'Sent' 1-6 denotes the sentence 7 was not recorded. The matching ratio refers to 'the number of the utterances whose target emotion and perceived emotion match' over 'the number of all utterances.' The matching ratios and associated standard deviations reflect the perceptual goodness of speech emotions portraited by the speakers.

Table 2 shows an example of confusion between target and perceived emotions. The evaluation result of one female evaluator on M1 is used for Table 2. A large number of 'other' implies that they are not pure target emotions, but possibly mixed emotions (e.g., happiness and sadness) and slightly different emotions (e.g., annoyed, fear, excited, nervous, worried) from the four target emotions. The USC-EMO-MRI corpus includes the evaluation results of every individual combination of speaker and evaluator.

Table 1: *Summary of evaluation results of all evaluators. '#Eval' denotes the number of evaluators. 'Sent' indicates the sentence ID included. 'AVE' and 'STD' denotes average and standard deviation of the matching ratio (%) between target emotion and evaluated emotion for sentence-level utterances, respectively.*

	Subject ID (M: male, F: female)									
	M1	M2	M3	M4	M5	F1	F2	F3	F4	F5
#Eval	10	10	11	10	11	12	12	12	12	10
Sent	1-6	1-7	1-7	1-7	1-7	1-6	1-6	1-6	1-7	1-7
AVE	85.3	69.5	82.6	72.0	80.5	80.7	94.0	89.8	86.5	80.5
STD	9.9	11.8	8.5	11.1	10.0	11.1	5.4	8.4	8.2	11.8

Table 2: *An example of confusion between the target emotion and the perceived emotion by a single evaluator for sentence-level utterance of M1's data.*

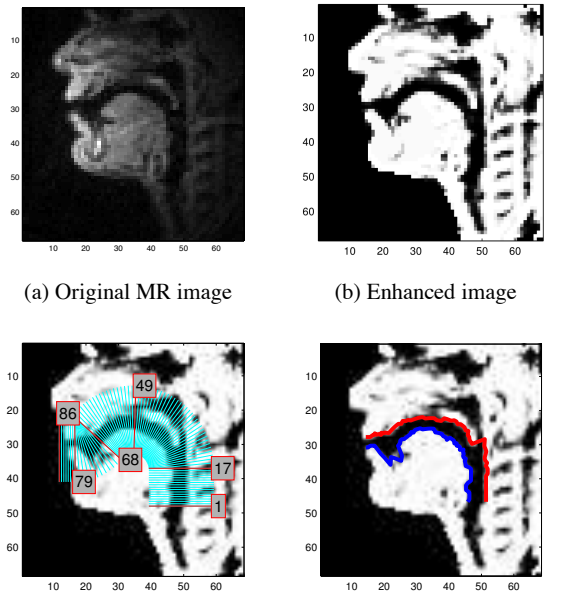
		Perceived				
		Neutrality	Anger	Happiness	Sadness	Other
Target	Neutrality	58	0	0	0	4
	Anger	2	42	1	0	7
	Happiness	1	0	38	0	13
	Sadness	2	0	0	39	11
	Total	63	42	39	39	35

4. Preliminary analysis of articulatory variation for different emotions

This section provides a preliminary analysis result on the MR images for emotion-dependent vocal tract movements. Vocal tract shapings of different emotions are compared in terms of cross-distances between the inner and outer vocal-tract walls from the larynx to the edge of the lips. In order to obtain these cross-distances, semi-automatic tissue-airway boundary segmentation was performed using a recently introduced MATLAB software [19]. This software performs (i) pixel sensitivity correction, (ii) noise suppression on the MR image, (iii) tracking of the lips and the larynx, (iv) segmentation of the airway-tissue boundary, and (v) computation of the distance function. The processes (iii) and (iv) are performed automatically based on the semi-automatically constructed gridlines. The cross-distances are obtained by measuring the Euclidean distance between the top-right and left-bottom airway-tissue boundary in each gridline from the larynx to the lips. The distance of the gridlines outside the upper airway (from the larynx to the lips) is considered to be 0. See [19] for the details of the method.

Fig. 2 illustrates the each step of parameter extraction process. Fig. 2 (a) shows an original MR image of the subject M1. Fig. 2 (b) shows the enhanced image by the pixel sensitivity correction and noise suppression. Fig. 2 (c) shows the gridlines. Finally, Fig. 2 (d) shows the tissue-airway boundary segmentation results.

Fig. 3 shows the mean of the range of the distance during the word 'five' in the sentence 1 for each emotion. In the plot, anger and happiness show wider movement range than sadness in the grid lines 49 ~ 68. The wider range of tongue movement for high arousal emotions (anger and happiness) in the palatal region than low arousal emotion (sadness) is well captured in the MR images. Fig. 3 also shows the difference between anger and happiness in terms of the tongue movement range; on average, anger shows wider movement range than happiness near the grid lines 13 and 23, while anger shows lower movement range than happiness in the grid lines 49 ~ 68. This indicates that on average, the pharyngeal constriction and releasing were more emphasized for anger than for happiness, while the palatal constriction and releasing were more emphasized for happiness than for anger, during producing 'five.'



(a) Original MR image (b) Enhanced image

(c) Gridlines (d) Tissue-boundary segments

Figure 2: *Vocal tract parameter extraction process. the grid lines are in cyan color. the number in the gray box denotes the gridline index. The red color line and the blue color line denote the right-top and left-bottom tissue-airway boundaries, respectively.*

5. Discussion and Future works

Although this corpus relies on the emotion elicited from actors/actresses in laboratory conditions, the emotional variability exhibiting in the data still allows us to study the controls over linguistic variables such as prosodic context, phonological environment and neutrality of utterances. Xu et al. also argues that "lab speech" is not necessarily inadequate for human speech production research [20].

We are also working on collecting EMA data from the same speakers using the same stimuli. The EMA data and the MRI data provide complementary information for articulatory dynamics. More specifically, EMA provides 3-dimensional sensor trajectories of certain articulatory surface points, while the collected rtMRI data provides image sequences of the full airway in the mid-sagittal plane represented by pixel intensity.

Development of robust (semi-)automatic MR image analysis tools for this corpus is another important on-going work of our research group. This task is important for improving the usability of this corpus in broader research communities, such as linguistics and pychology, with interests in speech production and perception.

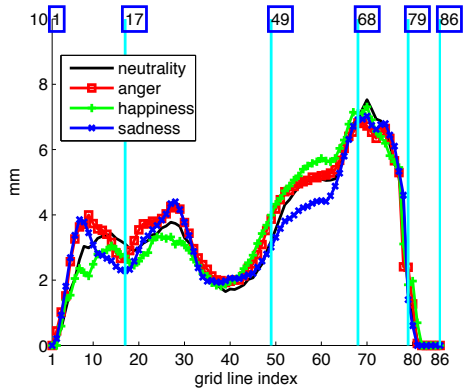


Figure 3: *The mean of the ranges of the cross-distance between estimated airway-tissue boundaries in each grid line for each emotion during uttering the word 'five.'*

The emotion evaluation for the USC-EMO-MRI corpus is still on-going process. To complement the emotion perception data from professional actors/actresses, the emotion evaluation from naive listeners will be also collected. Busso et al. [21] have reported the difference between professionals and naive listeners in terms of their emotion perception. It is interesting to investigate how the emotion perception of the two groups contrasts and to see whether the observation is a general trend in emotion perception.

6. Acknowledgements

This work is supported by NSF IIS-1116076 and NIH DC007124.

7. References

[1] D. Erickson, C. Menezes, and A. Fujino, "Some articulatory measurements of real sadness," Korea, 2004, pp. 1825 – 1828.

[2] D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya, "Exploratory study of some acoustic and articulatory characteristics of sad speech," *Phonetica*, vol. 63, no. 1, pp. 1 – 25, 2006.

[3] Donna Erickson, Osamu Fujimura, and Bryan Pardo, "Articulatory correlates of prosodic control: Emotion and emphasis," *Language and Speech*, vol. 41, no. 3-4, pp. 399–417, 1998.

[4] Jangwon Kim, Sungbok Lee, and Shrikanth S. Narayanan, "An exploratory study of the relations between perceived emotion strength and articulatory kinematics," in *Proceedings of Interspeech*, 2011, pp. 2961 – 2964.

[5] Jangwon Kim, Prasanta Ghosh, Sungbok Lee, and Shrikanth Narayanan, "A study of emotional information present in articulatory movements estimated using acoustic-to-articulatory inversion," in *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, 2012, pp. 1–4.

[6] Jangwon Kim, Sungbok Lee, and Shrikanth S Narayanan, "A study of interplay between articulatory movement and prosodic characteristics in emotional speech production.," in *Proceedings of Interspeech*. 2010, pp. 1173–1176, ISCA.

[7] Jeffrey J Berry, "Accuracy of the NDI Wave speech research system," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1295–1301, 2011.

[8] Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771 – 1776, 2004.

[9] Erik Bresch, Jon Nielsen, Krishna S. Nayak, and Shrikanth S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, Oct 2006.

[10] Sungbok Lee, Erik Bresch, Jason Adams, Abe Kazemzadeh, and Shrikanth S. Narayanan, "A study of emotional speech articulation using a fast magnetic resonance imaging technique," in *Proceedings of Interspeech*, Pittsburgh, PA, September 2006, pp. 2234 – 2237.

[11] Oudeyer Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 12, pp. 157 – 183, 2003, Applications of Affective Computing in Human-Computer Interaction.

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan 2001.

[13] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Theodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al., "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[14] Brad H Story, "Synergistic modes of vocal tract articulation for american english vowelsa)," *The Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3834–3859, 2005.

[15] Prasanta Kumar Ghosh and Shrikanth Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[16] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[17] Shrikanth Narayanan, Erik Bresch, Prasanta Kumar Ghosh, Louis Goldstein, Athanasios Katsamanis, Yoon Kim, Adam C Lammert, Michael I Proctor, Vikram Ramanarayanan, and Yinghua Zhu, "A multimodal real-time mri articulatory corpus for speech research.," in *Proceedings of Interspeech*, 2011, pp. 837 – 840.

[18] E. Bresch, Yoon-Chul Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]," *Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.

[19] Jangwon Kim, Naveen Kumar, Sungbok Lee, and Shrikanth Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," May 2014, (Accepted).

[20] Yi Xu, "In defense of lab speech," *Journal of Phonetics*, vol. 38, no. 3, pp. 329–336, 2010.

[21] Carlos Busso and Shrikanth S Narayanan, "The expression and perception of emotions: comparing assessments of self versus others.," in *Proceedings of Interspeech*, 2008, pp. 257–260.