AN ARTICULATORY STUDY OF EMOTIONAL SPEECH PRODUCTION

Sungbok Lee, Serdar Yildirim, 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

Few studies exist on the topic of emotion encoding in speech in the articulatory domain. In this report, we analyze articulatory data collected during simulated emotional speech production and investigate differences in speech articulation among four emotion types; neutral, anger, sadness and happiness. The movement data of the tongue tip, the jaw and the lower lip, along with speech, were obtained from a subject using an Electromagnetic articulography (EMA) system. The effectiveness of the articulatory parameters in emotion classification was also investigated. A general articulatory behavior observed was that emotionally elaborated speech production exhibits more peripheral articulations when compared to neutral speech. The tongue tip, jaw and lip positioning become more advanced when emotionally charged. This tendency was especially prominent for the tongue tip and jaw movements associated with sad speech. Angry speech was characterized by greater ranges of displacement and velocity, while it was opposite for sad speech. Happy speech was comparable in articulation to the neutral speech, but showed the widest range of pitch variation. It, however, remains to be seen if there is a trade-off between articulatory activity and voicing activity in emotional speech production. Multiple discriminant analysis showed that emotion is better classified in the articulatory domain. One probable reason is that the independency in the manipulation of each articulator may provide more degrees of freedom and less overlap in the articulatory parameter space. Analysis also showed distinct emotion effects for different phonemes: the high front vowel /IY/ was found to be less discriminated in both articulatory and acoustic domains than other peripheral vowels such as /AA/ and /UW/. It is likely that the physical boundary effect in the /IY/ articulation may leave less room to vary the tongue positioning and/or the lip configuration when compared to other vowels, resulting in less acoustic contrast among emotion types.

1. Introduction

We use emotions to express and communicate our feelings in everyday life. Our experience as speakers as well as listeners tells us that the interpretation of meaning or intention of a spoken utterance can be affected by the emotions that area expressed and felt. With recent advances in man-machine communication technologies, through automatic spoken dialog management systems, the question of how human emotion is encoded by a speaker and decoded by a listener has now attained practical importance. It is expected that by being able to detect emotion in user's speech and inject emotion into automatically generated system response depending on the dialogue situation, one can create a more natural man-machine interaction system. Such enhancements in man-machine communication technologies might help to draw more widespread use of such technologies.

It has long been known that speech prosody, that is, patterns in pitch and amplitude modulation and segmental durations including pauses, carry emotional information in the acoustic speech signal (c.f., [1]). Many studies have also made efforts to find acoustic correlates of emotions expressed in speech for automatic emotion detection [2] and emotional speech synthesis [3]. Although it is expected that underlying speech articulation, the major source of surface acoustics, might be affected by speaker's emotion, there has been only few published studies on emotional speech production. Those prior studies are also somewhat limited in scope in that the main focus has been on single speech articulators. For example, in [4], it was shown that the degree of jaw opening increases significantly as subjects become annoyed (or irritated) as they need to repeat the same answer in order to correct system response errors in a Wizard-of-Oz experimental setup. In [5], the lateral lip distance between the corners of the mouth is shown to be more influenced by emotion than by vowel identity itself. In [6], using EMA, it has been shown that the tongue positioning becomes more peripheral in sad emotion, especially for /IY/.

In this paper, we report some preliminary results on one fundamental aspect of emotion encoding by human, namely, "speech articulation" associated with emotional speech production. A specific focus is on vowel production. We recorded positions of three sensors attached to three major speech articulators, i.e., the tongue tip, the lower maxilla for jaw movement and the lower lip, using an EMA system while a subject produced simulated emotional speech. These data enable us to study both speech acoustics and underlying articulations as a function of emotion type. Such knowledge is valuable not only for the understanding of emotion encoding in the articulatory domain in conjunction from linguistic perspective but also for articulatory speech synthesis that can accommodate emotional coloring. In addition, from a theoretical standpoint, we examine the effectiveness of the articulatory parameters for emotion classification.

This paper is organized as follows: In the following section, the speech material and EMA data collection procedure are described. Data analysis procedures are described in section 3 and the results and findings are described in section 4. Discussion follows in section 5.

2. Data collection

2.1. Speech material

A set of 14 sentences (given below), which are mostly neutral in emotional content, was used for speech recording. A male native speaker of American English, who has no formal theatrical vocal training, 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 70 utterances in a row for each emotion resulting in a total of 280 utterances (14 sentences x 5 repetitions x 4 emotions). Each utterance was digitized in 12-bit amplitude resolution with 16kHz sampling rate. Speech was recorded simultaneously by the EMA system so that speech and corresponding articulatory movements are aligned in time.

The 14 sentences are: (1) I don't know how she could miss this opportunity; (2) Toby and George stole the game; (3) Your grandmother is on the phone; (4) They vetoed his proposal instantly; (5) Don't compare me to your father; (6) I hear the echo of voices and the sound of shoes; (7) Hold your breath and combine all the ingredients in a large bowl; (8) That dress looks like it comes from Asia; (9) They think the company and I will have a long future; (10) The doctor made the scar. Foam antiseptic didn't help; (11) That made being deaf tantamount to isolation; (12) The doctor made the scar foam with antiseptic; (13) I am talking about the same picture you showed me; (14) It's hard being very deaf. Tantamount to isolation.

2.2. EMA data acquisition

The Carstens' Ag200 EMA system was used to track the positions of three sensors in the midsagittal plane adhered to the tongue tip, the lower maxilla (for the 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 16kHz**.** Each sensor trajectory in the x-direction (forward-backward movement) and in the y-direction (vertical movement) with respect to the system coordinate is recorded by the EMA system.

After data collection, the raw articulatory data obtained by the EMA system were assembled into matlab data files. During the post-processing, each trajectory data was smoothed after correction for head movement and rotation to the occlusal plane so that the x-axis is parallel to the subject's occlusal plane. Finally, the origin of the coordinate system was translated to the upper maxilla reference position. Each sensor trajectory signal was then differentiated in order to obtain velocity components which were smoothed with a 9th order Butterworth filter of cutoff frequency 15-Hz**.**

It should be note that because of the convention of the coordinate system defined in this study, the forward movements of articulators (e.g., tongue tip advancement or lip protrusion) point to toward the negative x-axis (i.e., decreasing x coordinate) and the downward movements (e.g., widening tongue tip constriction or jaw opening) toward the negative y-axis (i.e., decreasing y coordinate). Articulatory behaviors and the axes shown in figures should be interpreted as such.

3. Data analysis

3.1. Acoustic analysis of speech

Each of the 280 utterances was first processed by an HMM-based forced-alignment procedure for automatic end pointing and phonetic segmentation. Then durations at the utterance and phoneme levels, including inter-word silence, were measured from the phonetically aligned speech signals. Pitch and the first three formant contours were also estimated for the individual vowel segments using the Praat speech processing software. Possible pitch and formant tracking errors were minimized using the procedures described in a previous study [7].

The duration, pitch and formant data were subject to statistical analysis, including analysis of variance (ANOVA) and multiple discriminant analysis, across phoneme and emotion variables using the SPSS statistical software package. For discriminant analysis of vowel class as a function of emotion type, minimum, maximum and median values of pitch and the first three formant frequencies as well as durations were used.

3.2. Articulatory analysis of EMA data

For each utterance, we measured the minimum, maximum, and average values of positions and velocities of the tongue tip, jaw and lower lip movements at the utterance level as well as the phonemic segmental level using the phonetically-aligned speech signals. Movements along the x and y axes were considered separately because they correspond to forward-backward and vertical movements, respectively, of each articulator. For vowel, they are each, respectively, correlated with the location and degree of the tongue constriction. For discriminant analysis of vowels across emotion types, minimum, maximum and average values of positions and velocities of three articulators in the x and y directions were used. Duration was also included.

4. Results

4.1. Duration and pitch

In Fig. 1(a) and Fig. 1(b), differences in durations among 14 sentences and averaged pitch values for vowels across emotional categories are plotted, respectively.

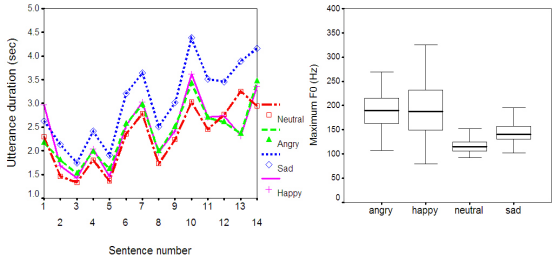

Figure 1. (a) Averaged sentence duration for each sentence. (b) Maximum pitch distribution.

On average, utterance durations become longer when speech is emotionally elaborated. ANOVA indicates that the effect of emotion is significant ($F(3,276)=93.1$, $p<0.00$). In Fig. 1(b) it is observed that elaborated emotional states are characterized by higher pitch and wider pitch distribution. The effect of emotions is significant ($F(3,2615)=278.9$,$p<0.00$). The largest degree of pitch modulation or pitch variability is associated with happy emotion. In fact, the wide pitch modulation seems the major means to simulate speech with happy emotion for the speaker.

4.2. Formant frequencies

The average first (F1) and second (F2) formant values of 4 peripheral vowels across emotions are plotted in Fig. 2. Differences in the F1 and F2 distributions of vowels suggest that the effects of emotion on vowel formants are different for different vowels. Specifically, it appears that other peripheral vowels are more influenced by an emotional change than the high front peripheral vowel /IY/. The same tendency has been observed in a previous study [7]. As can be seen later, this is likely due to articulatory constraints or saturation effect of tongue positioning associated with /IY/ articulation.

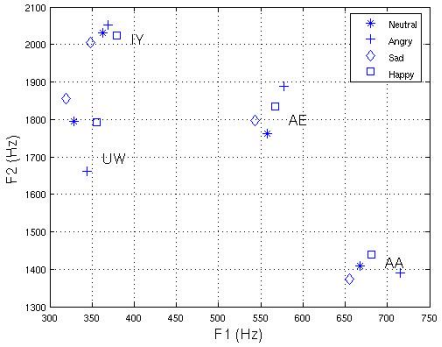

Figure 2. Averaged F1 and F2 values of 4 peripheral vowels for each emotion are plotted.

4.3. Articulatory kinematics

In Fig. 3, we use a phase space representation to visually illustrate the kinematic differences in articulatory movements as a function of emotion. Position and velocity of the tongue tip movement along the y-axis (i.e., vertical movement) for an instance of the productions of sentence #1 is shown in the figure. Differences in movement ranges and velocities among 4 emotions are clearly observable in the figure. Specifically, angry speech shows the largest movement range and velocity of the tongue tip vertical movement. In fact, this tendency for angry speech also holds for other articulatory movements such as jaw opening and tongue tip forwarding.

It is interesting to notice that the degree of tongue tip constriction (i.e., the closeness of the tongue tip to the roof of the mouth) increases for sad (bottom-left panel) and happy (bottom-right panel) speech. The increasing tongue constriction corresponds to the tongue tip movement toward the positive y-axis (i.e., toward the right side in each plot). This implies that the tongue tip articulation becomes more peripheral (or more upward in this case) when emotionally charged. This tendency seems a genuine phenomenon for the speaker examined in the study.

Results of a more detailed analysis for the 4 peripheral vowels (/IY/, /AE/, /AA/, /UW/) are given in Fig. 4 and Fig. 5. In Fig. 4, the averaged tongue tip x (top-left) and y (top-right) positions of each vowel are shown as a function of emotion and for the jaw at the bottom row. ANOVA indicates that most of the visual differences across vowels and emotions are significant (e.g., for the effect of emotion on the tongue tip x position, $F(3,512)= 35.7$, $p<0.00$).

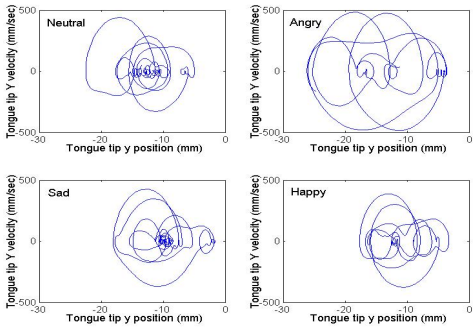

Figure 3. Phase space plot of the tongue tip movements along the y-axis (i.e., vertical movement). The utterance is of sentence #1.

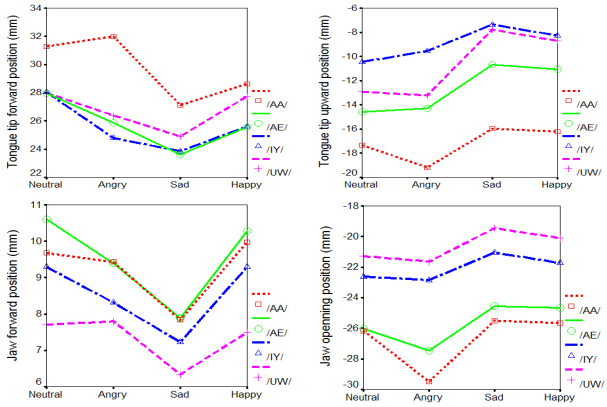

Figure 4. Averaged tongue tip x (top-left) and y (top-right) positions are shown for 4 peripheral vowels as a function of emotion. At the bottom, the case of jaw articulation is shown.

It is observed that the tongue tip exhibits most advanced and highest position for sad emotion and this tendency is universal for all the vowels, not only for /IY/ as reported in [6]. The same tendency also holds for jaw forward movement (bottom-left). As expected, the jaw opening is the largest for /AA/ and smallest for /UW/. For a given vowel, the jaw opening is the largest for angry emotion. This agrees with the previous observation in [4]. But such tendency seems universal for all the vowels, not only for /AA/, for angry emotion. It remains to be seen how this articulation associated with the angry speech is similar to irritated or emphatic speech articulations.

In Fig. 5, we show averaged tongue tip movement velocity of each vowel as a function of emotion. The effect of emotions is statistically significant. Angry speech shows the greatest tongue tip velocity and relatively smaller averaged velocity for sadness.

In summary speech articulation is most active for angry speech in terms of articulatory movements (i.e., displacement range and velocity) and sad is the least active. Based on these observations, one could characterize angry

speech as hyper-articulation and sad speech as hypo-articulation. It is interesting that happy speech articulation is more or less similar to neutral speech, except for the use of widest and highest pitch modulation by the speaker. It remains to be seen if there is a trade-off between articulatory activity and voicing activity in emotional speech production.

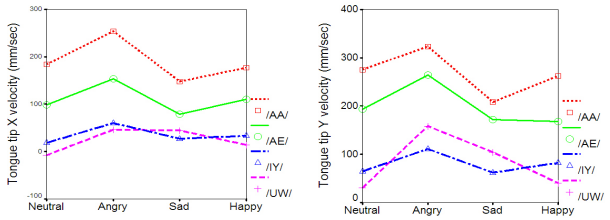

Figure 5. Tongue tip horizontal (left) and vertical (right) movement velocity plots of four peripheral vowels as a function of emotion.

4.4. Discriminant analysis

In Table 1, we shows the overall classification accuracy of ten monophthongal vowels using the multiple discriminant analysis.

	Neutral	Angry	Sad	Happy
Acoustic	81.4	54.4	60.4	55.8
Articulatory	69.7	59.0	73.8	77.5

Table 1. Overall classification accuracy (%) of vowel segments from 4 emotion categories. Duration is also included in the classifications.

In Table 2 and 3, classification accuracy of 4 peripheral vowels based on the acoustic and articulatory parameters are shown, respectively. In general, angry emotions show less accuracy and /IY/ shows the worst accuracy among vowels. Although the classification accuracy for /IY/ much improves in the articulatory domain, it still shows the lowest accuracy among vowels.

	Neutral	Angry	Sad	Happy
/IY/	73.6	50.0	67.6	64.1
/AE/	100.0	76.1	89.7	61.4
/AA/	95.6	75.6	81.8	64.4
/UW/	94.1	78.9	90.0	94.4

Table 2. Classification accuracy of 4 peripheral vowels based on acoustic parameters augmented with phonemic duration.

	Neutral	Angry	Sad	Happy
/IY/	100.0	86.9	95.6	96.6
/AE/	97.0	100.0	100.0	97.5
/AA/	100.0	100.0	100.0	100.0
/UW/	96.4	91.4	96.6	96.0

Table 3. Classification accuracy of 4 peripheral vowels based on articulatory parameters augmented with duration.

In summary, all the results indicate that emotions are more effectively discriminated in the articulatory domain. A probable reason is that the independency in the manipulation of each individual articulator may provide more degrees of freedom and less overlaps among them in the articulatory space.

It is also observed that the accuracy is better for the mid and back vowels than the front vowel /IY/ in both acoustic and articulatory domains. One probable reason seems that the saturation effect in /IY/ articulation may leave less room to vary the tongue positioning and possibly the lip configuration as a function of emotion. The saturation effect can be observed in the top-right plot in Fig. 4. It shows that the tongue-tip y position is the least variable for /IY/ as emotion varies. This might induce less acoustic variation of that vowel when compared to other peripheral vowels.

5. Discussion

Although the articulatory findings here can not be fully generalized because of the limitation in the number of subjects examined in the current study, we are still able to find some agreements with previous acoustic and articulatory studies and, extend them. For example, we observe that emotional speech articulation exhibits more peripheral or advanced tongue positioning, especially for sad emotion. This confirms the finding in [6]. In fact, such forward positioning of articulators is a general trend observed from the current subject. As shown in Fig. 5, the range of jaw opening is largest for the angry speech and this is in line with the finding in [4].

This study also indicates that the vowel /IY/ is less responsive to emotional changes when compared to other peripheral vowels. This illustrates the fact that the articulatory configuration associated with a vowel determines the effect of emotion on that vowel in the acoustic domain. As can be observed in Fig. 4 and 5, the effects of emotions on each articulatory parameter are fairly systematic across vowels. It is interesting to see if that observation is a general tendency in emotional speech production or just a speaker-dependent characteristic. It is noted that an informal listening test has been conducted by the first author to examine target emotions expressed by the speaker. This will be replaced by a formal listening experiment in order to fully rationalize the findings in this study. Those, and a detailed production study with data from more subjects, are topics of ongoing work.

Acknowledgements: The authors thank colleagues in the emotion research group for their valuable comments. This work was supported in part by grants from NIH and NSF.

6. 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] Bulut, M., Narayanan, S., and Syrdal, A., “Expressive Speech Synthesis Using a Concatenative Synthesizer,” ICSLP, Denver, 2002.

[4] D. Erickson, O. Fujimura, B. Pardo, “Articulatory correlates of prosodic control: Emotion and Emphasis,” Language and Speech, 41, 395-413, 1998.

[5] 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.

[6] D. Erickson, C. Menezes, A. Fujino, “Some articulatory measurements of real sadness,” ICSLP, Korea, 2004.

[7] 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. ICSLP, Korea, 2004