An Exploratory Study of the Relations between Perceived Emotion Strength and Articulatory Kinematics

Jangwon Kim[1], *Sungbok Lee*[1,2], *Shrikanth Narayanan*[1,2]

Signal Analysis and Interpretation Laboratory (SAIL)
[1]Department of Electrical Engineering, [2]Department of Linguistics
University of Southern California, USA

jangwon@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

Acoustic and articulatory behaviors underlying emotion strength perception are studied by analyzing acted emotional speech. Listeners evaluated emotion identity, strength and confidence. Parameters related to pitch, loudness and articulatory kinematics are associated with a 2-level (strong/weak) representation of the emotion strength. Two-class discriminant analyses show averaged leave-one-out accuracies of 65.8% and 63.8% in the acoustic and articulatory domains, respectively. Two-factor ANOVA (emotion type/strength) indicates that the listeners assess the emotion strength based on the nature of perceived emotions in the arousal dimension. Only hot anger and happiness show significant differences in pitch use in the strength contrast. Such contrasts are also observed in tongue lowering and/or advancing. The strength contrast by listeners may mainly rely upon pitch and loudness. However, interactions between the acoustic and articulatory parameters in strength perception are complex.

Index Terms: Emotional speech production, EMA, speech emotion, emotion perception, emotion strength.

1. Introduction

Emotions play an integral role in speech communication and affect the nature and course of the spoken dialog. Since the emotion expressed in speech is influenced by the listeners (interaction participants), it is desirable in emotions research to get the listener's assessment of not just the emotion category but also the perceived strength of the emotions. This present paper considers the relationship between the articulatory and acoustic correlates of the talker's emotional speech production and the perceived emotion type and strength. Such knowledge is valuable from both a theoretical standpoint (to shed further light on the production-perception link in speech communication) and application perspective (such as in informing better speech synthesis). In particular this work investigates the perceived strength of emotions based on human evaluations and attempts to relate those outcomes to the articulatory and acoustic properties of the (acted) emotional speech utterances that were evaluated.

A number of studies have documented the discriminative characteristics in prosodic and articulatory features for a particular emotion expression in speech. For example, statistics of pitch, in particular those at the utterance level have been widely shown to be effective for detecting emotions [1]. Likewise, the differences in articulatory position and velocity patterns across emotions have also been reported [2]. Despite the significant amount of previous work on the discriminative characteristics across categorical emotions, there are but a few studies that have considered the relationship between the strength of emotion expressed in an utterance and vocal effort, especially in terms of the underlying articulatory modulation. One such a study is [3], which showed that pitch variation is associated with emotional strength in perception. The main focus of current study is to investigate the kinematic articulatory behaviors to test the hypothesis that influence the perceived perceptual strength of emotions in speech. The corresponding acoustic correlates will also be investigated by analyzing pitch and intensity contours. Such comparisons would reveal the articulatory and acoustic contributions to the strength attribute of emotion perception. Those investigations will be based on the utterance-level statistics of prosodic and articulatory features.

This paper is organized as follows: First, the Electromagnetic articulography (EMA) data collection done for this study is described. Second, details of the listener evaluations process of the EMA data and their results are presented. Next, the prosodic and articulatory feature extraction and post-processing process are explained. The results and discussions of prosodic and articulatory characteristics for perceptual strength of emotion are followed. Finally, the summary of this study and directions for future work are provided.

2. Data Collection

2.1. Speech material

A set of seven sentences was spoken by a female (JR) and a male (SB) speakers. Both are native speakers of American English and have had theoretical vocal training. The set of sentences were repeated five times by the female speaker and four times by the male speaker in a random order. The speakers were asked to target one of the five emotional states and one of the three speech styles and produce each given utterance. The five target emotional states are neutrality hot anger, cold anger, happiness and sadness, and the three speech styles are normal, loud and fast. The list of seven sentences used for speech generation is following:

- Say peep again? That's wonderful.
- It was 9 1 5 2 8 9 5 7 6 2.
- Say pop again? That's wonderful.
- I saw 9 tight nightpipes in the sky last night.
- Don't know how very joyful he was yesterday.
- Say poop again? That's wonderful.
- Native animals were often captured and taken to the zoo.

2.2. EMA data acquisition

For each utterance, articulatory movements were recorded at 200 Hz sampling rate with an Electromagnetic articulography (EMA) system, simultaneously with speech waveform of a 16-kHz sampling rate. For the articulatory movements, the 3D-positions of 6 sensors (tongue tip, tongue body, tongue dorsum, upper lip, lower lip and jaw) were recorded with a Carstens' AG500 EMA system. The total number of utterances collected is 524 (7 sentences × 5 emotions × 3 styles × 5 repetitions - 1 bad recording) for JR and 417 (7 sentences × 5 emotions × 3 styles × 4 repetitions - 3 bad recording) for SB.

After recordings, head-movement corrections and occlusal plane correction of all utterances, the trajectory signal of each articulatory sensor was filtered with a 9th-order Butterworth filter with a 15 Hz cutoff frequency. Each sensor trajectory was also scanned for possible trajectory errors and erroneous trajectory segments were marked for exclusion during the analysis step.

3. Perceptual Evaluation of EMA Data

3.1. Emotion evaluation

Each utterance audio spoken by JR and SB was presented to five native American listeners (undergraduate or graduate students at the University of Southern California) in randomized order (e.g., http://sail.usc.edu/~jangwon/ema_eval_jr_short). The listeners for JR's data and SB's data are not identical. Listeners were asked to choose (1) the best-representative emotion among six emotion categories, such as neutrality, hot anger, cold anger, happiness, sadness and others, (2) confidence in their evaluation and (3) the strength of emotion expression. The listeners were asked to choose 'others' when they felt that none of the five given emotion categories best matched their perception. Confidence and strength were evaluated on a five-point scale.

3.2. Preprocessing of evaluation data

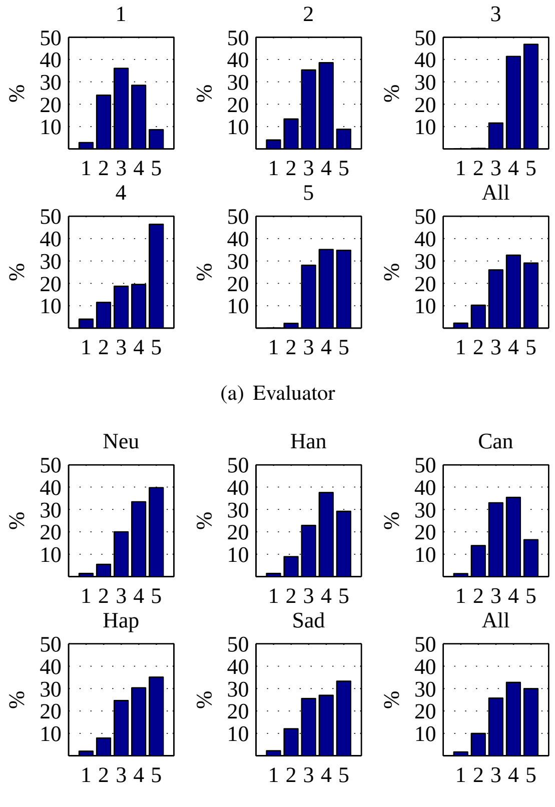
One of the evaluators for SB's data chose 'others' for 45.5% of utterances, even though the other evaluators chose 'others' for only 6.0% of utterances. Because of significantly different evaluation results, this evaluator's assessments were discarded from further analysis in this study.

The most representative emotion for each utterance was determined by majority voting (60% for JR's data and 50% for SB's data). If there were two emotions with same number of evaluations, then the one that obtained the higher confidence score was chosen. The confidence of each evaluator was z-scored along all utterance and used for fair comparison. An utterance was discarded if it does not satisfy the voting criteria.

Then, the final strength label of each utterance was determined as following. First, the average strength score of all evaluators was calculated for each utterance. Next, final strength label of an utterance was determined as weak if the average strength value was equal to or less than 40th quantile. It was determined as strong if the average strength value was equal to or greater than 60th quantile.

3.3. Emotion evaluation results and discussions

Figure 1 shows the histograms of raw strength scores depending on evaluators or emotions for JR's data (SB's data shows a similar trend). Overall, evaluations tend to be biased to high strength scores. It is also observed that the confidence score and strength score are highly correlated. The mean of correlation between confidence score and strength score of each evaluator across all emotions is 0.76 for JR's data and 0.72 for SB's data. It indicates that evaluators tended to be more confident of their evaluation when they felt strong expression than weak expression of emotion.



(a) Evaluator

(b) Emotion

Figure 1: *The histograms of strength scores in JR's data are presented. (a) is for each evaluator (1 to 5) and all evaluator (All). (b) is for each emotion and all emotion (All). Neu = neutrality, Han = hot anger, Can = cold anger, Hap = happiness, Sad = sadness*

Table 1 shows the confusion matrix between evaluated emotion (by evaluators) and target emotion (for speakers) used for analysis. A total 422 utterances of JR's data and 348 utterances of SB's data was used for analysis. The average identification rate between evaluated emotion and target emotion is 74.1% for weak expression and 87.1% for strong expression of JR's data, and 78.7% for the weak expression and 94.0% for strong expression of SB's data. Strong emotion expression matches better with the evaluated emotions. This is in line with the previously noted result that emotion expression strength is highly correlated with evaluation confidence.

There is much confusion especially between cold anger and hot anger in both JR's data and SB's data, which shows that speakers often failed to express the two emotions distinctively. Also, listeners chose hot anger significantly more than cold anger when they listened to loud style speech. As can be expected, loudness seems an important factor in recognizing hot anger more than cold anger. Happiness is also more biased to loud style speech than normal speech for strong emotion expression, while neutrality and sadness do not show significant

Table 1: *The confusion of evaluated emotion and target emotion for different strength (weak or strong) of emotion expression in perception. Neu = neutrality, Han = hot anger, Can = cold anger, Hap = happiness, Sad = sadness, Norm = normal.*

			Target emotion															
			JR								SB							
			Neu	Han	Can	Hap	Sad	Norm	Loud	Fast	Neu	Han	Can	Hap	Sad	Norm	Loud	Fast
Evaluated emotion	Weak	Neu	34	1	0	0	1	9	13	14	49	2	2	7	0	19	16	25
		Han	0	15	5	2	0	9	9	4	0	4	8	1	0	3	8	2
		Can	4	29	39	1	0	32	16	25	2	11	45	1	0	28	13	18
		Hap	0	0	0	20	3	5	7	11	0	1	0	27	0	10	6	12
		Sad	0	0	1	8	49	19	17	22	1	1	1	0	15	5	3	10
	Strong	Neu	51	0	0	0	0	20	14	17	16	1	0	1	0	13	4	1
		Han	0	31	12	1	0	8	23	13	0	47	6	1	0	2	25	27
		Can	0	13	17	0	0	3	9	18	0	1	7	0	0	3	3	2
		Hap	0	0	0	42	0	11	19	12	0	0	0	30	0	6	17	7
		Sad	0	0	0	1	42	21	15	7	0	0	0	0	56	20	26	10

bias to loudness. These results indicate that loudness may be a more important factor for high arousal emotions, such as hot anger and happiness, than low arousal emotions, such as neutrality, cold anger and sadness, in emotion strength perception. This results is consistent with previous studies of dimensional approach (activation-valence space) of emotion perception, e.g. [4]

4. Feature Measurements

4.1. Prosodic features

Statistics of fundamental frequency (in Hz) and intensity (in dB) of each utterance were used for analyzing the relation of prosody with perceptual strength of emotion expression. Fundamental frequency and intensity values were extracted in 10 milliseconds time step, using Praat [5]. After discarding pauses and silences, the pitch contours were examined for erroneous values using a 2-sigma criterion, and they were smoothed using a 5-point median filter. Absolute pitch derivatives (in Hz/second) were calculated from the cleaned and smoothed pitch contours. Intensity contours and intensity derivatives (in dB/second) were also smoothed in the same way as pitch contours.

After all the data pre-processing, we estimated the statistics of pitch, intensity and their derivatives in each utterance. We used quartiles, such as lower quartile, median, upper quartile and interquartile range in order to minimize the effect of still remaining noisy feature values.

4.2. Articulatory features

Articulatory trajectories were also pre-processed since they were sometimes noisy due to sensor-movement speed and direction related system errors. We checked the trajectory errors of all utterances manually and discarded articulatory measurements when reference sensor positions were not satisfactory after head movement correction. To remove this kind of error, we calculated the range of sum of the distances between all reference points. Then, all trajectories of an utterance were discarded if the sum is different from average by 1 millimeter. For errors of individual articulators, we noted the erroneous time regions and excluded them during analysis.

After smoothing each trajectory by a 5-point median filter, the position value (in millimeter) on the horizontal axis (x axis) and on the vertical axis (z axis) of each utterance were obtained. Velocity values (in millimeter/second) were also smoothed in the same way used for position values. Statistics of the articulatory measurements, such as x position, z position and tangential velocity for each utterance were used for analysis.

5. Results and Discussions

The significance of speech parameters for discriminating perceptual strength of emotion was tested by Fisher linear discriminant analysis using SPSS. Intra-speaker comparisons of utterance-level statistics are reported.

Table 2 and Table 3 show the classification accuracy of prosody statistics or articulatory statistics depending on the perceived strength of emotion expression. Leave-one-out classification implemented in SPSS was used. The baseline of classification accuracy (random case) is 50% (i.e., two-class).

Table 2: *Classification accuracy (%) of all prosody statistics or all articulatory statistics for perceptual strength of each emotion. Neu = neutrality, Han = hot anger, Can = cold anger, Hap = happiness, Sad = sadness, Pros: all prosody statistics, Arti: all articulatory statistics*

	JR					SB				
	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad
Pros	80.5	59.1	61.2	61.5	74.3	76.9	85.1	64.2	70.7	70.3
Arti	62.7	61.0	70.7	73.0	60.0	76.6	70.8	62.5	51.1	63.8

In Table 2, both prosody and articulatory parameters show considerable discriminant power for perceptual strength of emotions. The mean of classification accuracy of each emotion is 67.3% for prosody statistics and 65.4% for articulatory statistics in JR's data, 73.4% for prosody statistics and 65.0% for articulatory statistics in SB's data. The averaged classification accuracies (across two subjects) on all emotions data are 65.8% and 63.8% in the acoustic and articulatory domains, respectively. An interesting observation is that there are many cases (neutrality, cold anger, happiness and sadness of JR, and happiness and sadness of SB) of relatively high classification accuracy in either prosody statistics or articulatory statistics. So, it is speculated that there are complex interactions between the acoustic and articulatory domains in the perceptual strength contrasts by listeners. Hot anger of JR and cold anger of SB show relatively low classification accuracy in both prosody and

Table 3: *Classification accuracy (%) of subsets of articulatory statistics for each emotion. Neu = neutrality, Han = hot anger, Can = cold anger, Hap = happiness, Sad = sadness, TT: tongue tip, TB: tongue body, TD: tongue dorsum, L: lips, J: jaw, Tp: tongue position, Tv: tongue velocity, Lp: lip position, Lv: lip velocity, Jp: jaw position, Jv: jaw velocity. The highest accuracy in each box is highlighted.)*

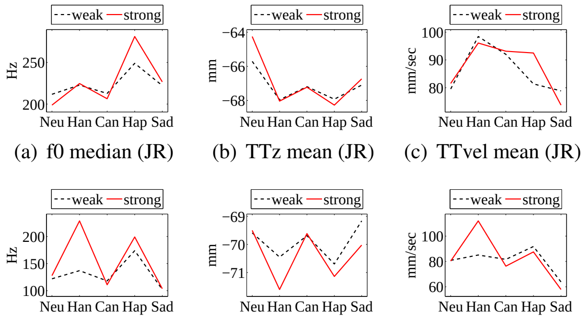
	JR					SB				
	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad
TT	63.2	55.7	67.0	73.0	69.0	74.0	74.1	54.7	51.9	71.0
TB	71.1	62.3	60.2	48.6	62.5	75.3	73.2	68.3	57.4	72.1
TD	64.5	61.0	58.8	70.3	62.5	70.0	65.4	46.8	55.6	59.0
L	72.0	54.1	55.1	51.4	62.5	57.6	68.5	64.4	54.8	71.7
J	65.8	59.0	51.1	70.3	62.0	68.5	79.3	68.8	51.9	75.8
Tp	73.7	62.7	59.0	73.0	59.2	77.1	66.0	45.9	57.8	53.3
Tv	67.1	49.2	59.0	64.9	64.8	68.6	80.0	63.9	55.6	61.7
Lp	64.0	63.9	57.3	54.1	51.4	56.1	72.2	67.8	34.0	50.0
Lv	52.0	65.6	40.4	62.2	62.5	30.3	72.2	44.1	49.1	75.0
Jp	64.5	60.7	55.7	64.9	62.0	68.5	79.3	65.6	53.7	61.3
Jv	53.9	55.7	53.4	67.9	60.6	61.6	79.3	50.0	53.7	75.8

articulatory statistics. It may imply that in these cases perceptually important factors for emotion strength contrast do not lie in the current feature dimensions.

In Table 3, the mean of the highest classification accuracies in each box is 69.7% in JR's data and 71.9% in SB's data. This result also supports the hypothesis that articulatory modulations are significantly associated with perceptual strength of emotion expression.

Two-factor (emotion type and strength) ANOVA analyses with individual statistics show that in general the listeners assess the strength attribute based on the nature of perceived emotions in the arousal dimension, both in the articulatory and acoustic domains. For instance, hot anger and/or happiness have shown significant differences in pitch in the weak and strong contrasts by listeners [e.g., F=7.68, p<0.01 for happiness in JR's data, F=32.09, p<0.01 for hot anger in SB's data]. Such a contrast is also observed in some statistics of articulatory movements, for example the tongue tip (TT) lowering, advancing and/or TT velocity [e.g., F=7.03, p=0.01 for TT vertical position range for hot anger in SB's data, F=6.94, p=0.01 for TT horizontal position median for hot anger in SB's data, F=8.66, p<0.01 for TT velocity maximum of happiness in JR's data]. Some example plots of prosody and articulatory statistics for strength contrast of each emotion are presented in Figure 2.

Lastly, it is observed that significant parameters (grouped or individual) for perceptual strength contrast of each emotion are not consistent across speakers in many cases. For example, in Table 2 prosody statistics of hot anger in SB's data show the highest classification accuracy among five emotions, while they show the lowest classification accuracy in JR's data. It indicates that speakers have different modulation schemes which are associated with perceptual strength of emotion expression. Also, it seems that the perceptual strength of emotion might not be related to any single speech production parameters, but rather it is associated with combined modulations in both the acoustic and articulatory domains.



(a) f0 median (JR) (b) TTz mean (JR) (c) TTvel mean (JR)

(d) f0 median (SB) (e) TTz mean (SB) (f) TTvel mean (SB)

Figure 2: *Example plots of prosody and articulatory statistics. TTz = tongue tip position on the vertical axis (z-axis), TTvel = tongue tip tangential velocity, Neu = neutrality, Han = hot anger, Can = cold anger, Hap = happiness, Sad = sadness.*

6. Conclusions

In this paper, we investigated the relations of perceptual strength of acted emotion expression in relation with articulatory movements as well as prosodic characteristics. Our analysis was based on the pitch and articulatory statistics at the utterance level. Even though broadly generalizable conclusions cannot be drawn from this initial study due to its limited number of subjects and data samples, this study found evidence supporting that articulatory movements are significantly associated with perceptual strength of emotion in general. First, averaged classification accuracies using articulatory statistics by Fisher linear discriminant analysis is 63.8%. Also, it was observed that there are statistically significant articulatory parameters related to perceptual strength of emotion, especially for high arousal emotions, shown by two-way ANOVA analyses. Acoustic attributes such as pitch and loudness were also shown to be highly associated with perceptual strength of emotion expression.

It is speculated that the strength contrast by listeners may rely upon the acoustic attributes such as pitch and loudness. However, there seem to exist complex interactions between the acoustic and articulatory domains in cuing the strength contrast by listeners. Inter-speaker difference is another major factor. Investigating these questions are part of on-going research work in our lab.

7. References

[1] Busso, C., Lee, S. and Narayanan, S., "Analysis of emotionally salient aspects of fundamental frequency for emotion detection", IEEE Transactions on Audio, Speech and Language Processing, 17(4):582-596, May 2009

[2] Lee, S., Yildirim, S., Kazemzadeh, A. and Narayanan, S., "An Articulatory study of emotional speech production", Interspeech, Lisbon, Portugal, pp. 497-500, 2005

[3] Scherer, K. R., Ladd, D. R., Silverman, K. E. A., " Vocal cues to speaker affect: Testing two models", Journal of the Acoustical Society of American, 76, 1346-1356, 1984

[4] Laukka, P., Juslin, P., Bresin, R., "A dimensional approach to vocal expression of emotion", Cognition and Emotion, Volume 19, Number 5, 633-653(21), August 2005

[5] Boersma, P., Weenink, D., "Praat: doing phonetics by computer (Version 5.1.08)", Retrieved May 11, 2009, from http://www.praat.org/