A Detailed Study of Word-Position Effects on Emotion Expression in Speech

Jangwon Kim, Sungbok Lee and Shrikanth S. Narayanan

Viterbi School of Engineering, University of Southern California, Los Angeles, California, USA
jangwon@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

We investigate emotional effects on articulatory-acoustic speech characteristics with respect to word location within a sentence. We examined the hypothesis that emotional effect will vary based on word position by first examining articulatory features manually extracted from Electromagnetic articulography data. Initial articulatory data analyses indicated that the emotional effects on sentence medial words are significantly stronger than on initial words. To verify that observation further, we expanded our hypothesis testing to include both acoustic and articulatory data, and a consideration of an expanded set of words from different locations. Results suggest that emotional effects are generally more significant on sentence medial words than sentence initial and final words. This finding suggests that word location needs to be considered as a factor in emotional speech processing.

Index Terms: Emotion, Speech Analysis, Emotion recognition, Acoustic features, Articulatory features

1. Introduction

In recent years, several research efforts on emotional speech analysis have been carried out. A widely adopted approach to examining the emotional effects on speech is to use acoustic signal features, and it has been shown that emotion influences variations in spectro-temporal properties of segmental and suprasegmental (prosodic) parameters [1]. For example, the mean values and standard deviation of Mel Frequency Cepstrum Coefficients (MFCCs) and Mel Filter Bank (MFBs) have been shown to manifest significant emotional cues, and are in fact widely used as feature for automatic emotion recognition [2], [3]. Other signal features that carry significant emotional information include fundamental frequency, energy, formants, speech rate and ratio of duration between stressed and unstressed region.

Another venue for studying the effects of emotion is to examine articulatory movements. While only few studies have followed this line of research, mainly due to the difficulties in acquiring direct articulatory data, specific differences have been detected when emotion is conveyed in speech. For example, the articulatory positioning in vocal tract during emotional speech is more extreme than that of neutral speech [4], [5]. In addition, basic emotions, such as neutral, angry, sad and happy, were intended by speakers, articulator position, range and velocity were modulated but within the constraint that articulators' movement for reaching linguistic targets points were not compromised [6].

It was also reported that the vertical range of Jaw opening is significantly correlated with the some prosodic change of acoustic signal in [7]. These approaches can inform the analysis proposed in this work.

The heterogeneous nature of spoken language signal, and the underlying generation mechanisms are well known. One motivating reason for the present research is to explore how emotional effects co-vary given the heterogeneous nature of the unfolding spoken utterance. For example, do different parts of an utterance display different effects of emotion? A literature survey indicates that there has been no explicit study on the emotional expression as function of word position in a sentence. Understanding the variation patterns of emotional effects in word locations could be useful for improving the schemes of automatic recognition, for example.

The general goal of this study is to examine our hypothesis that emotional effects manifested in the speech signal depend on the word position within a sentence. To examine this, we collected articulatory and acoustic data about the same word which was located at different positions in sentences and observed the variation patterns of emotional effects depending on the word location.

This paper is organized as follows. First, datasets, feature extraction and emotion modeling methods are described in section 2. Next, the results and discussions of emotion localization in each experiment are provided in section 3. Finally, the summary of this study and directions for future work are provided in section 4.

2. Method

2.1. EMA database

The articulatory data used for this study comes from Electromagnetic articulography (EMA) database, which contains a total of 680 utterances spoken by three native speakers of American English, two females and one male [4]. Only one female talker had previous training in theatre/acting. The two female talkers produced 10 sentences, and the male produced 14 sentences, 10 of which are common sentences with those of the female talkers'. Each sentence was repeated 5 times for each of the four different emotions: neutral, angry, sad and happy.

This database has both speech waveforms at 16 kHz sampling rate and synchronized articulatory movement measurements at 200 Hz sampling rate. Articulatory movement measurements include position, velocity and acceleration values of the tongue tip (TT), lower lip (LL) and jaw (Jw) on x and y axis. By forced-alignment using HMMs, articulatory movement measurements were first automatically aligned to corresponding phonetic segments at the word level. After that, the alignments were checked, and corrected if needed, manually.

2.2. Articulation labeling

To see the emotional effects along different word positions, we chose the word "TANTAMOUNT" for our initial analyses. This word is included in two sentences, "It's hard being very deaf. Tantamount to isolation." and "That made being deaf tantamount to isolation." This word was selected not only because it is located at both phrase initial position and

sentence medial position, but it also requires relatively a large number of articulatory gestures to be produced.

For labeling, we marked 7 points within the region of the selected word by hand. A sample is shown in Figure 1. The 7 points indicate the first TT closure point (ttcl1) for /t/, TT's lowest point in /t ae n/ region (ttlow1) for /ae/, the second TT closure point (ttcl2) for the second /t/, the first LL closure point (llcl1) for /m/, the third TT closure point (ttcl3) for the last /t/, the point of maximum velocity of TT in /t ae n/ region (velmax) and the point of minimum velocity of TT in /t ae n/ region (velmin). For this labeling, the values on only y axis were considered. Label points were decided where each corresponding articulator reaches 97% out of the range from the nearest proceeded target position to the present target position.

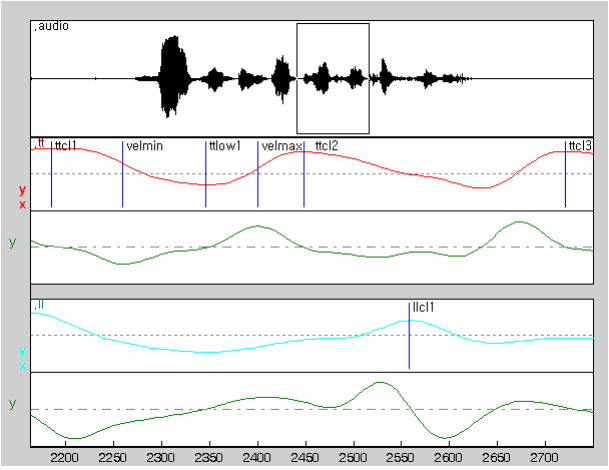

Figure 1: *Labeling of 7closure or releasing time of TT and LL, such as ttcl1, velmin, ttlow1, velmax, ttcl2, llcl1 and ttcl3, based on the trajectory of the articulators on y axis. Subplot of '.tt' illustrates TT position and velocity; subplot of '.ll' illustrates LL's. Upper panel for each articulator indicates position measurement and lower panel indicates velocity.*

Of the 120 utterances (3 speakers x 4 emotions x 2 sentences x 5 repetitions), 6 utterances including abnormal first TT closure movements were excluded from this analysis. The abnormal movements had occurred because some speakers tended to begin the TT's approaching motion to the next target point early when pause existed between phrase and sentence. Another utterance whose data file was corrupted in the experiment was also excluded. In summary, the total number of utterances used for this analysis is 112.

2.3. Feature extraction from the manual labels

To examine how the degree of emotional context changes depending on word location, we chose 9 articulation features. The list of the features used is provided in Table 1. The feature set includes various articulatory information, including duration of articulatory gestures, articulatory phase, maximum speed of articulator motion and the range of articulatory movement. Articulatory phase refers to the relationship of gestures between different articulators. Since articulatory gestures and articulatory phase are influenced by articulatory velocity and positioning, we assumed that emotion can also affect variation in these features. For duration normalization, all duration features except total duration are divided by total duration. All duration features except total duration as well as duration between LL closure (LLCL) and TT closure (TTCL) indicate duration ratios within stressed syllable region, /t ae n/. Duration between LLCL and TTCL indicates duration ratio of non-stressed syllable region, /ta/. Duration between TT maximum speed points indicates the duration ratio of the interval between TT's maximum velocity point and minimum velocity point. Low velocity and High velocity indicate the minimum velocity value and maximum velocity value of TT motion, respectively. Lastly, Y Range indicates vertical TT movement range in /t ae n/ region.

For the analysis of these features, we used the SPSS statistical software package for analysis of variance (ANOVA). The four basic emotional speech data are categorized to binary emotion groups: emotional speech (angry, sad and happy) and neutral speech (neutral).

2.4. Data selection for acoustic and articulatory analysis

To improve the generalizability of the analysis, we next expanded the examination of word position-dependent emotional effects to both articulatory domain and acoustic domain with more testing words. All utterances in EMA database were evaluated by four native listeners based on which emotion they belonged to. The degree of emotional strength on each utterance was recorded by numeric score over 0 to 5. Based on this evaluation, 174 utterances which received poor emotion score were excluded in this study. The 4 sentences produced by only the male speaker were also excluded. Therefore, a total 406 utterances were used for expanded acoustic and articulatory analysis.

These 406 utterances were separated into training dataset and testing dataset. Basically, testing dataset has feature vectors in the region of testing words, such as FOAM, TANTAMOUNT, DEAF, SCAR and ANTISEPTIC. FOAM and TANTAMOUNT were chosen for the comparison between phrase initial word and sentence medial word; DEAF, SCAR and ANTISEPTIC were used for comparison between sentence medial and final words. The training dataset was composed of the feature vectors extracted from 301 utterances which do not include the testing words. In general, stressed speech segment has different acoustic properties from non-stressed syllables, e.g. higher pitch and energy. For this reason, we examined two types of segments for TANTAMOUNT and ANTISEPTIC, the whole word segment and stressed syllable segment, respectively. The starting and ending points of all segments were obtained from labels in the corresponding word file, which includes syllable, starting point and ending point.

2.5. Acoustic and articulatory movement feature extraction

For acoustic analysis of emotion localization, two sets of features were used. One set includes 12th-order MFCCs, normalized energy and their delta values; the other set includes 12th-order MFBs, normalized energy and their delta values. All these features were extracted from wave files using HTK tool. The speech acoustic signal was bandlimited between 0 and 4 kHz. High-frequency pre-emphasis was applied before the feature extraction procedure.

We assumed that average syllable length is about 80-100 millisecond, and used 80 milliseconds as window length and 10 milliseconds for window shift time. From the EMA database, we directly extracted position and velocity values on x axis and y axis, and their tangent vectors of all three articulators, TT, LL and Jw. It is true that the duration for fast TT motions, like flaps, are less than 10ms. However, this articulatory feature setting is enough to capture the motion

because it includes maximum, mean and variance values of position and speed measurements in a window. Identical window and shift time with acoustic features' are used. The total number of initial features for articulatory analysis is, therefore, 54 (3 positions and 3 velocities x 3 articulators x mean, maximum and variance).

We also assumed that some of the features in the articulatory domain are correlated. For example, the movements of Jw and LL are highly correlated. In this case, Principle Component Analysis can be used in order to provide only uncorrelated feature sets. By this process, computation complexity can be reduced. In this study, we reduced the number of articulatory features from 54 to 10 by the threshold 90% of the total variance being explained by the reduced set.

2.6. Training and testing of emotion models by Gaussian Mixture

For acoustic and articulatory analysis, each of the two target classes, emotional speech and neutral speech, was modeled by Gaussian Mixture Models (GMMs). The frames of silent regions at the beginning and the end of utterances were excluded. Each model was trained by using the expectation-maximization algorithm. To find the best mixture settings, we empirically increased the number of mixtures and tested the resulting performances. The number of mixture for the final model was decided when the average emotion recognition rate for utterances increased less than 1 %. As a result, 8 mixtures were chosen for acoustic analysis. The average emotion recognition rate of GMMs for MFCC features was 84.08%; recognition rate for MFB features was 85.37%. In this case, the recognition rate of MFB features was slightly higher than that of MFCC, but this tendency was not consistent in different window length settings. Each frame in non-silence regions of testing dataset was classified based on log-likelihood ratio. After that, the class of each utterance was decided by majority voting.

The 2 mixture models were chosen for the binary emotion models in articulatory domain. The mixture selection followed the same criterion for acoustic emotion models. The average recognition rate by the models was 98.37% for all utterances of testing dataset. The reason for the better performance of articulatory model than acoustic models may be due to the fact that original feature dimension for articulatory feature is considerably higher than acoustic feature dimension, which may reflect the emotional state better.

2.7. Testing emotional effects on different word locations in acoustic and articulatory domain

We analyzed emotional effects on different word locations with the average emotion recognition rate of each word position. Using Gaussian classifier, the frames in a word segment in an utterance were classified. The class of each word segment is decided by majority voting. After that, the emotion recognition rates of each word location were obtained.

3. Result and Discussion

3.1. Emotion localization on articulatory features for the word "TANTAMOUNT"

The results of two-way ANOVA are shown in Table 1. The significance (sig.) values of some features are not as clear as we expected. In general, articulation for sad speech is closer to neutral speech than happy and angry speech in terms of speech rate and articulation velocity [4]. This tendency seems to move the mean values of emotional features close to those of neutral, which can reduce sig. values. It was shown by one-way ANOVA for sad speech samples vs. angry and happy speech samples. In the results for medial position, all duration features except the duration between LLCL and TTCL, the duration between TT maximum speed points and High velocity were significant (<.05). For initial position, only velocities and total duration were significant (<.01). These results confirm that the different deviation patterns of sad from other emotions blur the significance, especially of the duration of TT releasing (TTRL) in medial position and total duration in initial position.

For all duration features in stressed syllable region in TANTAMOUNT, such as dur. of TTRL, dur. of TTCL and dur. between 2 TTCLs, of sentence medial position are clearly more significant than those of phrase initial position. This result suggests that the emotion reflected in stressed syllable causes stronger deviation of articulatory gestures on sentence medial position than phrase initial position. The high significance value of the duration between LLCL and TTCL on both word positions indicates that emotion does not cause significant variations on the articulatory phase between TT gesture and LL gesture, presumably for maintaining linguistic integrity.

Table 1: *Two-way ANOVA analysis results. Duration (Dur.) of TTRL: (ttlow1 – ttcl1)/(ttcl3 – ttcl1); Dur. of TTCL: (ttcl2 – ttlow1) /(ttcl3 – ttcl1); Dur. between 2 TTCLs: (ttcl2 – ttcl1)/(ttcl3 – ttcl1); Dur. between LLCL & TTCL: (llcl1 – ttcl2) /(ttcl3 – ttcl1); Total dur.: (ttcl3 – ttcl1); Dur. between TT max. speed points: (velmax – velmin)/(ttcl3 – ttcl1). Low velocity: velocity at velmin point; High velocity: velocity at velmax point; Y Range: greatest value of TT position between ttcl1 and ttcl2 – TT position at ttlow1. Significant results are shown by highlight.*

Feature	Initial position		Medial position	
	F value	Sig.	F value	Sig.
Dur. of TTRL	.10	.96	2.35	.08
Dur. of TTCL	.57	.64	3.66	.02
Dur. between 2 TTCLs	.67	.58	4.15	.01
Dur. between LLCL & TTCL	.78	.51	.53	.67
Total dur.	2.84	.05	4.91	.01
Dur. between TT max. speed points	1.06	.37	2.19	.10
Low velocity	2.64	.06	2.31	.09
High velocity	2.54	.07	2.21	.10
Y Range	2.75	.05	4.92	.05

In another perspective, dur. of TTRL, dur. of TTCL and dur. between 2 TTCLs are relatively highly related to TT motion and are significant only at sentence medial position. Duration between TT max. speed points, which is related to TT motion, is also more significant on sentence medial position than phrase initial position. It may indicate that TT gestures are highly correlated with emotion in stressed region.

Low velocity, High velocity, Total duration and Y Range are significant for both phrase initial position and sentence medial position. It may explain that the articulation speed and vertical movement range of TT reflect emotion well, but position-dependent emotional effects are not apparent.

3.2. Emotion localization in acoustic domain

In this section, the emotion localization effects are examined in acoustic feature domain. Table 2 shows the difference between phrase initial position and sentence medial position; Table 3 shows the difference between sentence medial position and final position. The results in Table 2 and Table 3 indicate the mean values of recognition rates from emotional speech and neutral speech testing data.

Table 2: *Emotion recognition rates of sentence initial position vs. sentence medial position. TAN is TANTAMOUNT; ANT is ANTISEPTIC; S indicates only stressed syllable of word. Significantly higher recognition rates are shown by highlight.*

		FOAM	TAN	TAN S
MFCC	Initial	.61	.88	.88
	Medial	.62	.86	.86
MFB	Initial	.64	.88	.89
	Medial	.89	.86	.86

Table 3: *Emotion recognition rates of sentence medial position vs. sentence final position. TAN is TANTAMOUNT; ANT is ANTISEPTIC; S indicates only stressed syllable of word. Significantly higher recognition rates are shown by highlight.*

		DEAF	SCAR	ANT	ANT S
MFCC	Medial	.78	.77	.84	.85
	Final	.58	.67	.85	.82
MFB	Medial	.85	.91	.86	.86
	Final	.67	.89	.88	.84

In these two tables, we can observe some general trends. First, the emotional effects on sentence medial words are more significant than sentence final words. Also, the emotion recognition performance of MFBs is mostly better than MFCCs. This result is similar to those obtained in the previous study in [3]. In addition, emotion localization effects are more clearly detected on single-syllable words, such as FOAM, DEAF and SCAR, than multi-syllable words, such as TANTAMOUNT and ANTISEPTIC. In general, multi-syllable words have more chance to convey emotional information than single-syllable words because of longer word duration. In this sense, speakers' intended emotion is expressed well on multi-syllable word regardless of its location.

3.3. Emotion localization in articulatory domain

As a final step, we examined the emotion localization with articulatory movement features. Table 4 shows the mean value of emotion recognition rates from emotional speech and neutral speech testing data as in Tables 2 and 3.

Table 4: *Emotion recognition rates of speech segments in different positions when articulatory movement features are used. TAN is TANTAMOUNT; ANT is ANTISEPTIC; S indicates only stressed syllable of word. Significantly higher recognition rates are shown by highlight.*

	FOAM	TAN	TAN S
Initial	.99	.93	.93
Medial	1.00	1.00	1.00

A: Sentence initial position vs. sentence medial position.

	DEAF	SCAR	ANT	ANT S
Medial	1.00	1.00	.99	.99
Final	.95	.99	.97	.97

B: Sentence medial position vs. sentence final position.

In Table 4, the overall recognition rates are high for all segments (>0.93). As a result, strong emotion localization differences as seen in the acoustic domain cannot be expected. However, we can make some interesting points based on this table. For single syllable words, sentence medial position is more significant than other positions in all cases. This tendency was also shown in acoustic domain. It confirms that emotion localization may be detected well at single syllable words which do not have enough emotion information. Another interesting point is that emotional effects on sentence medial words are stronger than phrase initial words. It is similar to the result of articulatory gesture features. Therefore, it confirms that emotion localization between phrase initial words and sentence medial words is reflected in the articulatory domain as well.

4. Conclusions

In this paper, we investigated the emotion localization at the word level in both acoustic domain and articulatory domain. The summary of this study is listed below.

In general, the emotional effects on sentence medial words are more significant than phrase initial words and sentence final words. Emotion localization on single-syllable words is ubiquitous in both acoustic and articulatory domain. Also, binary emotion models based on MFBs reflect emotion localization better than MFCCs in word level.

For this initial emotion localization study, all indexes were labeled manually. An automatic segmentation is necessary for future studies to allow us process larger amounts of data to generalize the emotion localization effects on articulatory gesture features. In addition, the number of words located at more than one location was limited. Therefore, a more elaborate database is also needed.

In conventional emotional speech studies, the speech segments are chosen based on several criteria. For example, common nouns or their phonetic segments are commonly used for emotional speech study, not pronouns. Our study indicates that word location within an utterance is another criterion to be considered in emotional speech processing.

5. References

[1] Lee, C. and Narayanan, S. "Emotion recognition using a data-driven fuzzy inference system," Eurospeech, Geneva, Switzerland, pp. 157-160, 2003.

[2] Grimm, M., Mower, E., Kroschel, K. and Narayanan, S., "Primitives-based emotion and evaluation of emotions in speech" Speech Communication, 49: 787-800, 2007.

[3] Busso, C., Lee, S. and Narayanan, S., "Using Neutral Speech Models for Emotional Speech Analysis," Interspeech, Antwerp, Belgium, pp. 2225-2228, 2007.

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

[5] Erickson, D., Menezes, C. and Fujino, A., "Some Articulatory Measurements of Real Sadness," ICSLP, Korea, pp. 1825-1828, 2004.

[6] Lee, S., Bresch, E. and Narayanan, S., "An Exploratory Study of Emotional Speech Production using Fundamental Data Analysis Techniques" 7th International Seminar On Speech Production, Ubatuba, Brazil, pp. 525-532, 2006.

[7] Erickson, D., Fujimura, O. and Pardo, B., "Articulatory Correlates of Prosodic Control: Emotion and Emphasis" Language and Speech, pp. 395-413, 1998.