# A Study of Interplay between Articulatory Movement and Prosodic Characteristics in Emotional Speech Production

*Jangwon Kim*[1], *Sungbok Lee*[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

This paper investigates the interplay between articulatory movement and voice source activity as a function of emotions in speech production. Our hypothesis is that humans use different modulation methods in which articulatory movements and prosodic modulations are differently weighted across different emotions. This hypothesis was examined by joint analysis of the two domains, using two statistical representations: (1) the sample distribution comparison using two-sigma ellipses of the articulatory speed statistics and prosodic feature (pitch or intensity) statistics, (2) the comparison of correlation coefficients. In the articulatory-prosodic spaces, we find (1) distinctive weighting patterns for angry and happy emotional speech and (2) distinctive correlation patterns depending on articulators and target emotions. These findings support the hypothesis that humans use different modulation methods of emphasizing articulatory motions and/or prosodic activities depending on emotion.
**Index Terms**: interplay, articulatory control, prosodic control, emotional speech production, joint analysis

## 1. Introduction

Traditionally, emotional speech production research has been conducted under the "source-filter" assumption that the production mechanism is composed of two components – the source activities and the vocal tract shaping – which are combined linearly. Fundamental frequency (pitch) and intensity are the most representative parameters for describing voice source behavior, while articulatory movement range and speed have been used to characterize vocal tract shaping patterns. Speech modulation methods for delivering emotional information have been studied based on these features, for example in [1, 2, 3].

Previous studies have reported that pitch statistics are critical prosodic cues, providing discriminative power for distinguishing emotions e.g., [4]. Intensity was also reported as an important parameter for vocal emotion expression, notably in the arousal (activation) dimension [5, 6]. However, there are only few studies in the articulatory domain due to the difficulties of direct articulatory data collection. Distinctive patterns of articulatory motions for emotion expression were studied in [2, 7]. These studies have discussed the distinctive variation patterns in aspects of prosodic and articulatory features.

Despite the progress in understanding critical prosodic and articulatory features for target emotion encodings, the knowledge of how humans "jointly" control the features in both the prosodic domain and the articulatory domain is still limited. This study explores the interplay between articulatory motion and voice source behavior through pitch and intensity variations, as a function of emotion in speech. We believe that understanding how voice source signals and articulations co-vary emotionally can yield new insights in the broad mechanisms of human speech production. Such knowledge can lead to, for example, more human-like articulatory synthesis with emotion coloring.

Our hypothesis is that humans use different modulation methods in combining the prosodic and articulatory aspects depending on the target emotion. For example, speakers may put more weight on the articulatory domain controls or on prosodic domain controls depending on distinctive emotion encoding goals. This hypothesis is examined in this work by sample distribution comparison and analysis based on the statistics of prosodic and articulatory features. We present results of speaker-independent characteristics as well as speaker-dependent characteristics used for emotion encoding. We also present results about some constraints on the joint controls in the prosodic and articulatory domains.

The remainder of this paper is organized as follows: In Section 2, we describe the database, feature extraction and data analysis methods. In Section 3, we describe the experimental results and findings with discussions. Finally, conclusions and future work description follow in Section 4.

## 2. Methods

### 2.1. Experimental setup

We used the electromagnetic articulography (EMA) database, collected at the University of Southern California [2]. This database includes speech waveforms and the corresponding articulatory movement measurements of three articulators, tongue tip (TT), lower lip (LL) and jaw (JAW). During the data collection, 10 sentences were repeated five times for each of four emotion states, anger, happiness, neutrality and sadness, by three speakers. The three speakers are a male (AB) and two females (JN and LS). JN had vocal training in acting, and her happy speech is somewhat more exaggerated than that of the other speakers. Articulatory motion was recorded as x (forward-backward movement) and y (vertical movement) coordinates for each articulator at 200 Hz sampling rate. The speech waveforms were collected at 16kHz sampling rate simultaneously with the articulatory data. Phonetic labels were created by forced alignment using HMMs first, and then they were manually corrected.

In the dataset of 600 utterances (10 sentences $\times$ 5 repetitions $\times$ 4 emotions $\times$ 3 speakers), only 503 utterances which contain clear emotional content, judged so by human evaluators,

were included in this study. The last word segment region in an utterance was excluded, because its prosodic and articulatory behaviors generally have some variations, including low pitch, low intensity and slow articulatory motions. Lastly, be-verbs and function words, such as prepositions, articles, pronouns and particles, were excluded, because emotional information tends to be encoded more in other parts of speech, such as nouns, verbs and adjectives [8].

Finally, we chose consonant-vowel (CV) and vowel-consonant (VC) segments which have (closure + releasing) motion or (approaching + closure) motion of either TT or LL. The segments which include a diphthong, a fricative or a vowel before a retroflex were excluded. Table 1 shows the list of CV and VC segments included in this study.

Table 1: The list of CV and VC, present in the DB and included in this study. CRA indicates the critical articulator for consonant production. Upper cases in the word column indicate the tested syllable region. In the stress column, 1 indicates stressed, 0 indicates a single-syllable word, -1 indicates unstressed.

| C | V | C | Word | CRA | Stress |
|---|---|---|---|---|---|
| | AE | N | grANdmother | TT | 1 |
| M | AH | | grandMOther | LL | -1 |
| D | AA | | DOctor | TT | 1 |
| | AX | M | cOMpare | LL | -1 |
| B | IY | | BEing | LL | 0 |
| D | EH | | DEAf | TT | 0 |
| T | AE | | TAntamount | TT | 1 |
| | AE | N | tANtamount | TT | 1 |
| T | AX | | tanTAmount | TT | -1 |
| | AX | M | tantAMount | LL | -1 |
| | AH | M | cOMpany | LL | 1 |
| P | AX | | comPAny | LL | -1 |
| | AX | N | compANy | TT | -1 |
| N | IY | | compaNY | TT | -1 |
| L | AO | | LOng | TT | -1 |
| | AE | N | ANtiseptic | TT | 1 |
| T | AX | | anTIseptic | TT | -1 |
| | EH | PD | antisEPtic | LL | 1 |
| T | IX | | antisepTIc | TT | -1 |
| T | AO | | TAlking | TT | 0 |
| P | IH | | PIcture | LL | 1 |
| L | UH | | LOOks | TT | 0 |
| | AH | M | cOMes | LL | 0 |

### 2.2. Feature extraction

The pitch and intensity contours of each utterance were extracted using the Praat speech processing tool [9], which allows various parameter settings for pitch extraction. We applied different parameter settings for each emotion and each speaker. Pitch and intensity were extracted with a 250 milliseconds window shifting by 50 milliseconds. The extracted raw pitch and intensity values of each vowel region in CV or VC were manually checked under the assumption that the pitch values of normal speech do not change by more than 50 Hz within 50 milliseconds, and the segments including error values in the vowel region were excluded from pitch data samples. Next, the pitch and intensity contours were smoothed by a 5-point median filter and then by an 8th order butterworth low pass filter with 25 Hz cutoff frequency. Lastly, we excluded the segments in which the number of extracted pitch values is less than the half of the number of frames in its vowel segment. This happened where pitch estimation is difficult, for example, the breathy voice regions and laryngealized vowel regions. The final pitch and intensity values in the vowel region of each CV or VC segment were used for analysis.

Tangential articulatory speed was calculated by using the articulatory velocity values of TT, LL and JAW in the EMA database. The start and the end times of consonants and vowels were extracted from the phonetic labels in the database. We used articulatory speed values instead of articulatory position values for analysis in order to minimize the variations depending on vowels. We used the CV and VC regions instead of just the vowel regions, because the critical articulator's releasing or approaching motions in the transition regions are not always included within only vowel regions. For example, an extended aspiration time of the preceding voiceless plosive consonant reduces the duration of the vowel region, resulting in that releasing motion is not included within the region.

### 2.3. Data analysis

We used the feature statistics, maximum and range to analyze the interrelation between prosodic features and articulatory features. In this paper, we report only the results of maximum pitch, maximum intensity and maximum articulatory speed, because the difference among the results with other statistics were not significant. Maximum pitch and maximum intensity were calculated in the vowel region of each CV or VC segment, while the maximum articulatory speed of its critical articulator was calculated in the whole region of CV or VC. Correlation coefficients and p-values were calculated based on the maximum articulatory speed values and the maximum pitch values, or the maximum articulatory speed values and the maximum intensity values. The analysis was based on the sample distribution plots, correlation coefficients and p-values for the four emotions.

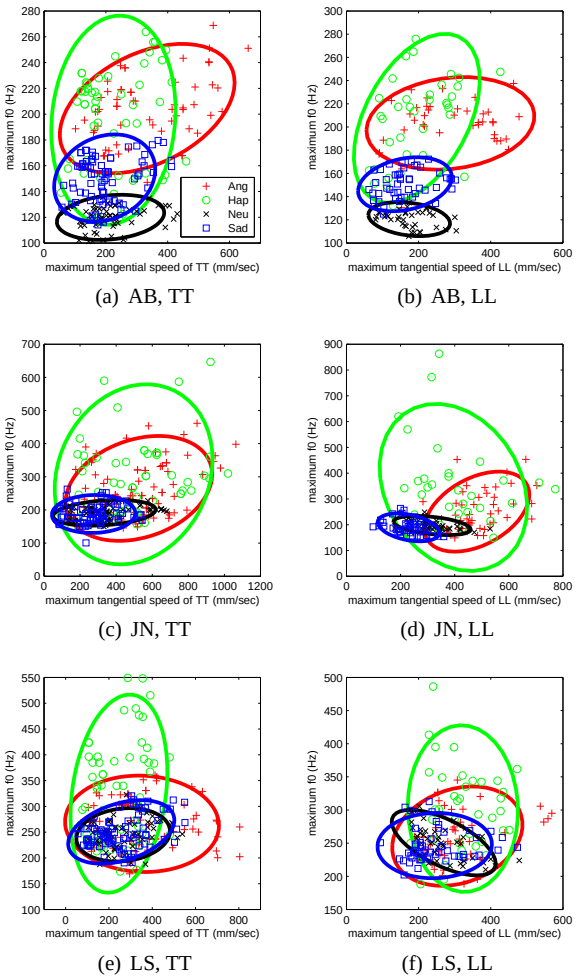
## 3. Results and Discussions

This section will discuss the interplay between the articulatory domain and the prosodic domain during emotion expression based on the results in Fig. 1, Fig. 2 and Table 2. The plots related to JAW are omitted in this paper due to their high similarity with those of LL in terms of sample distribution shapes, presumably because LL motions are highly correlated with JAW motions.

### 3.1. Interplay between pitch and articulation

The plots of maximum articulatory speed and maximum pitch for four emotions can be seen in Fig. 1. Some speaker-independent patterns are observed in Fig. 1. The sample distributions of happy speech show higher variations in the maximum pitch (f0) dimension than the maximum articulatory speed dimension. However, those of angry speech shows higher variations in the maximum articulatory speed dimension than the maximum f0 dimension. These tendencies indicate that the speakers emphasize articulatory speed modulation for expressing anger, while they emphasize pitch modulation for expressing happiness. These patterns are observed on the plots of JAW as well. Also, sad speech shows slightly higher variations on the maximum pitch dimension than neutral speech, even though their variation size is relatively smaller than those of angry speech and happy speech.

In neutral speech, all plots of LL and JAW for all speakers




(a) AB, TT (b) AB, LL

(c) JN, TT (d) JN, LL

(e) LS, TT (f) LS, LL

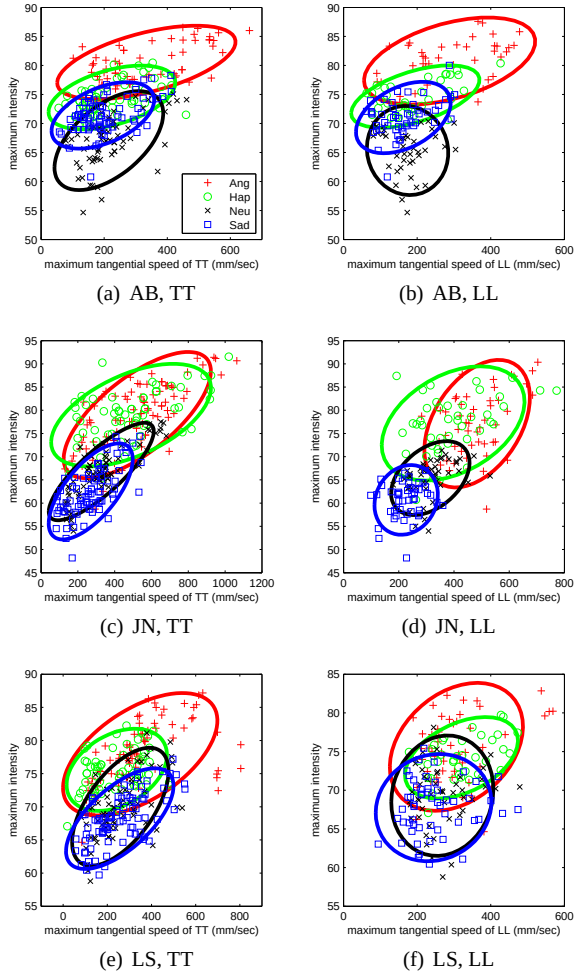
Figure 1: Example plots of the maximum tangential speed of critical articulators and the maximum pitch. A circle indicates the Gaussian contour with 2 sigma standard deviation for each emotion (red-Ang, green-Hap, black-Neutral, blue-Sad). Different emotions show distinctive variation patterns in the articulatory speed dimension and the pitch dimension (see details in text).

show (even if slightly) negative correlation between the maximum articulatory speed and the maximum pitch. The underlying reason may be that there are different modulation patterns between the stressed region and the unstressed regions (which are linguistically determined). Compared to the unstressed region, the sample distributions of the stressed region showed greater variations in the maximum articulatory speed and lower variations in the maximum intensity. This tendency implies that, for stress in neutral speech, vocal controls are concentrated on the articulatory speed modulations.

### 3.2. Interplay between voice intensity and articulation

The plots of maximum TT or LL speed and maximum intensity for the four emotions can be seen in Fig. 2. The correlation coefficient values and p-values for maximum speed of each articulator and maximum intensity are reported in Table 2. A significant speaker-independent tendency observed in Fig. 2 is the positive correlation between maximum articulatory speed and maximum intensity for angry and happy. These results indicate that high-arousal emotion expression may accompany significantly correlated controls of articulatory speed and intensity. In fact, this positive relationship was universally observed in all other plots except the case of LS's JAW. In the case of LS's JAW, such relationship disappeared, which is reflected in Table 2 ($p = 0.76$ for anger, $p = 0.57$ for happiness), probably due to her idiosyncratic vocal characteristics.

The positive correlation between maximum articulatory speed and maximum intensity were more significant for TT ($p \leq 0.005$) than LL and JAW (Table 2, Fig. 2). This observation indicates that intensity modulations are more correlated with TT speed modulations than LL and JAW speed modulations.




(a) AB, TT (b) AB, LL

(c) JN, TT (d) JN, LL

(e) LS, TT (f) LS, LL

Figure 2: Example plots of the maximum tangential speed of critical articulators and the maximum intensity. A circle indicates the Gaussian contour with 2 sigma standard deviation for each emotion (red-Ang, green-Hap, black-Neutral, blue-Sad). Different emotions show distinctive variation patterns in the articulatory speed dimension and the intensity dimension (see details in text).

Table 2: The correlation values and p-values (in parenthesis) of the maximum intensity and the maximum speed of critical articulators. Strong positive correlation between articulatory speed and intensity is shown for angry and happy, and TT. (ARTI = critical articulator, Ang = Anger, Hap = Happiness, Neu = Neutrality, Sad = Sadness)

| ARTI | Emotion | AB | JN | LS |
|---|---|---|---|---|
| TT | Ang | 0.56(0.00) | 0.69(0.00) | 0.55(0.00) |
| | Hap | 0.46(0.00) | 0.53(0.00) | 0.37(0.00) |
| | Neu | 0.55(0.00) | 0.83(0.00) | 0.60(0.00) |
| | Sad | 0.47(0.00) | 0.67(0.00) | 0.65(0.00) |
| LL | Ang | 0.45(0.00) | 0.41(0.00) | 0.32(0.02) |
| | Hap | 0.54(0.00) | 0.38(0.01) | 0.42(0.01) |
| | Neu | -0.05(0.74) | 0.42(0.00) | 0.11(0.47) |
| | Sad | 0.41(0.00) | 0.15(0.29) | 0.10(0.48) |
| JAW | Ang | 0.45(0.00) | 0.42(0.00) | 0.04(0.76) |
| | Hap | 0.43(0.01) | 0.36(0.02) | 0.09(0.57) |
| | Neu | -0.25(0.12) | 0.00(0.98) | -0.15(0.32) |
| | Sad | 0.46(0.00) | -0.17(0.26) | -0.36(0.01) |

### 3.3. Speaker-dependent characteristics

AB shows some unique patterns for sad speech, compared to the other speakers. Firstly, AB shows more significant correlation ($p \leq 0.005$ in Table 2) and a strong positive relationship between the maximum articulatory speed and the maximum intensity for sad speech. This relationship was not as strong as neutral speech (Fig. 2). Secondly, AB's maximum pitch values and maximum intensity values of sad speech are greater than those of neutral speech (Fig. 1, Fig. 2), while the sample distributions of neutral speech and sad speech are not significantly different for JN and LS. These results suggest that AB emphasizes pitch controls and intensity controls more than articulatory controls for expressing sadness distinctively from neutrality. JN and LS may emphasize other dimensional controls, for example voice quality controls, which was reported as another important control dimension for sadness in [10].

Another interesting result is the interrelated vocal instrument modulations for exaggerated happiness in JN's speech. Her sample distribution plots show significantly higher variations in both the maximum f0 dimension and the maximum articulatory speed dimension with reduced correlation between the two dimensions, while also showing significantly higher variations and high correlation patterns with angry speech in the maximum intensity dimension and the maximum articulatory speed dimension. These results indicate that JN additionally used more articulatory speed modulations almost independently with pitch modulations, but maintaining the correlation with intensity modulations for exaggerated happiness. We need to collect data from additional subjects to further validate these findings.

## 4. Conclusions and Future Works

In this paper, we investigated the interplay between certain prosodic prosodic features and articulatory movement during emotion expression in speech. Our analysis was based on the pitch and intensity statistics for prosody and the statistics of tangential speed of critical articulators in CV and VC segments. Even though broadly generalizable conclusions cannot be drawn from this small data set, this study detected some significant patterns consistently shown across speakers. First, in the articulatory speed-pitch statistics space, speakers tend to emphasize articulatory speed modulations for angry speech, while emphasizing pitch modulations for happy speech. Secondly, a significant positive correlation between intensity statistics and articulatory speed statistics is observed for high-arousal emotions (anger and happiness). It is also found that the correlation are significant for all emotions in the TT speed-intensity space.

The articulatory information used in this study is limited to TT, LL and JAW from three speakers, so we are planning to collect more detailed articulatory data (including real time MRI) for more detailed experiments. Also, more emotional speech data including richer vocal modulations will be useful to validate and further investigate some of the interesting findings, like exaggeration in speech. Future work also includes the development of emotional speech production models that account for the interplay between the prosodic factors and articulatory factors. These models can inform the design of more realistic articulatory synthesis that improve the perceived naturalness of emotional speech production.

## 5. Acknowledgement

Thanks to colleagues in the SAIL lab for their comments.

## 6. References

[1] Lee, C. M., Narayanan, S. S., "Toward detecting emotions in spoken dialogs", in IEEE Transactions on Speech and Audio Processing, 13:2(293-303), 2009

[2] Lee. S., Yildirim, S., Kazemzadeh A., Narayanan, S. S., "An articulatory study of emotional speech production", in Proceedings of InterSpeech, pages 497-500, 2005

[3] Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S. S., "An acoustic study of emotions expressed in speech", in Proceedings of InterSpeech, pages 2193-2196, 2004

[4] Busso, C., Lee, S., Narayanan, S. S., "An analysis of emotionally salient aspects of fundamental frequency for emotion detection", in IEEE Transactions on Audio, Speech, and Language Processing, 7:4(582-596), 2009

[5] Banse, R., Scherer, K. R., "Acoustic Profiles in Vocal Emotion Expression", J. Pers. Soc. Psy.,vol. 70(3), p. 614-636, 1996

[6] Kienast, M., Sendlmeier,W., "Acoustical analysis of spectral and temporal changes in emotional speech", in ISCA Workshop on Speech and Emotion, Northern Ireland, 2000

[7] Erickson, D., Menezes, C., Fujino, A., "Some articulatory measurements of real sadness", in Proceedings of Interspeech, pages 1825-1828, Korea, 2004.

[8] Bulut, M., Lee, S., Narayanan, S. S., "A statistical approach for modeling prosody features using POS tags for emotional speech synthesis", in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 1237-1240, Honolulu, Hawaii, 2007

[9] P. Boersma, "Praat, a system for doing phonetics by computer", Glot International, vol. 5, no. 9/10, pp. 341-345, 2001.

[10] Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T., Shibuya, Y., "Exploratory study of some acoustic and articulatory characteristics of sad speech", in Phonetica 63:1-25, 2006