Relations between prominence and articulatory-prosodic cues in emotional speech

Jangwon Kim, Anil Ramakrishna, Sungbok Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory
Viterbi School of Engineering, Univ. of Southern California, Los Angeles, CA
http://sail.usc.edu

Abstract

This study investigates the relations between the degree of prominence and articulatory-prosodic cues in emotional speech. In particular, this study considers articulatory parameters driven from the Converter/Distributor (C/D) model. The goal is to obtain a better understanding of the link among syllable magnitude in the C/D model, the empirical way to measure it in literature, and syllable-level prominence, and to examine emotional variations appearing in this relation. Since prosodic variations are important cues for prominence and emotion in speech, relations with prosodic parameters (f0, energy, duration) are also considered. Electromagnetic articulography data of two speakers were used for analysis. The degree of prominence was computed on crowd-sourcing annotation data, using the Rapid Prosody Transcription. Results indicate that movements of linguistically critical articulator, energy, syllable magnitude measure are highly correlated with prominence; f0 is relatively less correlated. The movements of linguistically critical articulator tend to be more correlated than syllable magnitude measure. Inter-speaker variability and emotion-dependent variations are also reported. These results suggest complex relations between prominence and articulatory-prosodic cues. They also suggest that incorporating more articulatory and prosodic behaviors than the conventional way can better relate to perception of prominence.

Index Terms: Converter/Distributor model, prominence, emotional speech, prosody, speech production, Rapid Prosody Transcription

1. Introduction

Prominence refers to perceptual quantity of standing out from its environment [1]. Speech prominence is an important paralinguistic cue in natural verbal communication, because it carries paralinguistic and key linguistic messages in speech utterance. This is also crucial for decoding semantic or pragmatic focus, lexical stress or boundaries [2–5]. Acoustic variations imposed on prominent speech segments (e.g., syllables and words) are created by controls of voice source and articulatory movements. This study investigates the relationship between prominence and articulatory-prosodic behaviors, and revisits the conventional way of representing syllable-level prominence in the Converter/Distributor (C/D) model proposed by Fujimura [6].

The C/D model is a comprehensive speech production model that algorithmically relates the metrical structure of speech utterance and surface-level speech signals, such as articulatory movements, f0 and energy. Syllable is the minimal unit of the metrical structure in this model. The prosodic characteristics of individual syllable are represented as a function of syllable magnitude (i.e., prominence) and syllable duration. Theoretically speaking, syllable magnitude and syllable duration represent *inherent* prosodic characteristics of the corresponding syllable in an abstract space. Previous studies have proposed algorithmic ways of computing the two parameters from vertical trajectories of articulators [6], the validity of which have also been examined in literature [7, 8]. The conventional approach to computing syllable magnitude is to measure Jaw excursion within the corresponding syllable region [**?**, 6, 8, 9], under the assumption that the degree of jaw opening is well correlated to the perceptual strength of the corresponding syllable [10, 11].

However, syllable magnitude influences the movements of linguistically critical articulator, as well as prosodic behaviors indirectly according to the C/D model theory. In fact, relying on jaw excursion itself is not optimal for emotional speech, production of which often involves various jaw controls, e.g., clenched jaw for cold anger [8]. The main goal of this study is to examine statistical relationship between articulatory-prosodic parameters and the degree of prominence, and inform the C/D model of a better algorithmic way of representing the syllable magnitude for emotional speech. The C/D model is appealing for modeling and analysis of speech production, because it offers a comprehensive theoretical framework of speech production from speech planning stage, to execution and realization stages. In particular, this modeling theory offers simple and intuitive ways of representing prosodic characteristics of speech utterance, which is attractive for emotional speech production modeling. Novel findings in this study can be useful for better implementation of the C/D model in terms of the relationship among different stages of speech production.

There are several ways of determining the degree of prominence of speech segments in an utterance in literature. For knowledge-based approach, discrete prominence score is assigned based on metrical-phonological rules [12]. The rules for determining prominence score is summarized well in [13]. For data-driven approach (i.e. using subjective prominence ratings), several methods in different scales have been proposed in previous studies [14, 15]. However, direct annotation of the relative value is not reliable and consistent without professional training [5]. An alternative approach is to quantify perceptual prominence based on simple binary annotation from multiple ordinary evaluators [16]. The present study employs the latter proposed by Cole [17], which represents *group's perception* obtained from multiple ordinary listeners. This method allows us to easily represent the degree of prominence based on the manual annotation. Results in previous studies indicate that this representation of prominence is reliable and consistent, as well as comparable to the ones obtained by linguistic experts [18]. The details of this method is provided in Section 2.1.

This paper is organized as follows: Section 2 describes the data and feature extraction. Section 3 reports our results and discussion. Finally, Section 4 offers our conclusions and plans for future work.

2. Methods

2.1. Data

We used ElectroMagnetic Articulography (EMA) data and simultaneously recorded speech audio. This data was recorded while one male and one female native speakers of American English read two sentence prompts. The speakers had professional vocal training for theatric performance. Before reading the prompts for each emotion, the speakers were asked to immerse themselves into the target emotion. After immersing themselves to each of five target emotions, the speakers uttered each sentence five times in randomized order. The target emotions are neutral, angry, happy, fearful and sad.

The two sentence prompts were designed for studying C/D model parameters. The sentence prompt 1 is *"Pam said bat that fat cat at that mat,"* and the sentence prompt 2 is *"Nine one five, two six nine, five one six two."* It must be noted that sentence prompt 1 was developed in our previous study [8] for studying the behaviors of the C/D model parameters. Eight out of nine words in total contain identical vowel /AE/ so that the variations of Jaw movements due to vowels is minimized. Also, eight words contain stop or fricative consonants at coda and onset positions of each syllable, where articulatory parameters of the C/D model can be algorithmically computed. For sentence prompt 2, six words out of 10 in total contain stop or fricative consonants at coda and onset positions of each syllable. For the syllables without onset or coda consonant in both sentences, articulatory gestures of preceding coda or following onset consonant were considered, but carefully by manual inspection.

We followed the methods of our previous study [8] for data acquisition and processing, and emotion quality evaluation. The present paper offers summary of these methods (See [8] for more details). Both articulatory data and speech waveform were recorded in parallel, using the NDI WAVE system and a directional microphone. Since we are interested in the movements of linguistically critical articulator (for onset and coda) and the jaw, we used the EMA data of the tongue tip, the tongue dorsum, the lower lip, and the jaw. The critical articulator was determined based on the place of articulation. The six Degree Of Freedom (DOF) sensor in the NDI system was used as the reference point, while the 5-DOF sensors were used to monitor the anatomical points for the interested articulators. Speech waveform was initially recorded at a sampling rate of 22050 Hz, while EMA data was recorded at a frame rate of 100 Hz. The recorded EMA data was rotated based on the occlusal plane sensor data (three 5-DOF sensors) so that the sensor positions were located on the true occlusal plane of each speaker. Then, the individual articulatory sensor trajectories were smoothed using a Butterworth low pass filter at a cut-off frequency of 20 Hz. The emotion quality of each utterance was evaluated by at least 11 native speakers of American English. The final emotion label for each utterance was determined by using majority voting criterion.

The degree of word-level prominence is represented using p-score (prominence score) [17]. The p-score is computed as an average of binary prominence evaluation (prominent: 1 vs not-prominent: 0) of 20 naive listeners; hence it is a floating point number in [0, 1]. This evaluation system is based on the Rapid Prosody Transcription (RPT) [17] which is a fast way of determining binary prominence classes (prominent or not-prominent) for speech segments. Each sentence prompt was displayed to the listeners while the listeners were asked to mark prominent word(s) after listening to the corresponding speech audio. All listeners are ordinary English speakers in the United States. The anootations were collected on Amazon Mechanical Turk.

2.2. Feature extraction

We extracted syllable-level (i.e., monosyllabic-word-level) prosodic and articulatory features. Prosodic features comprise acoustic duration and various statistics of Root-Mean-Squared (RMS) energy and f0 (pitch). The acoustic duration for each word was obtained by subtracting the starting time from the ending time of the word, computed based on acoustic forced alignment. We used a hidden Markov model based forced aligner, called SailAlign [19], and manually corrected the alignment outputs afterwards. The RMS energy and f0 were computed with a 25 msec window and 10 msec shifting. Articulatory features driven from the C/D model comprise syllable magnitude and inter-iceberg (time) distance. Syllable magnitude has been computed conventionally by measuring jaw excursion within the corresponding syllable. Inter-iceberg distance is the time duration between the iceberg points for the onset and the coda consonants of the corresponding syllable. See [8] for details of the algorithm for computing the two C/D model parameters. The other articulatory features comprise various statistics of the vertical positions of (i) linguistically critical articulator for the onset and (ii) the jaw. The statistics comprises mean, maximum and range of raw frame-level feature values and its derivative for all frame-level features (RMS energy, the vertical positions of the critical articulator and the jaw), except f0. We used median, 75% quantile and interquartile range for f0 and its derivative in order to minimize the effect of noise in extracted f0 features. In total, 27 articulatory-prosodic features were extracted.

3. Results and Discussion

3.1. Prominence depending on speakers and emotions

Prior to analyzing the relations between prominence and articulatory-acoustic cues, we examined whether our prominence score (p-score) varies significantly depending on speakers and emotions. To analyze differences in prominence scores between the groups (i.e., speaker and emotion), we performed two-way Analysis of Variance (ANOVA) with respect to speaker ID and emotion type. Results of ANOVA revealed statistically significant difference among the five emotions (F-statistic = 6.61; p-value $< 10^{-4}$). Figure 1 shows boxplots of p-scores (prominence score) depending on emotions. The p-score tends to be lower when the listeners perceived low arousal emotion (sad) than when they perceived high arousal emotions (angry and happy). We examined this pattern using one-sided t-test on two groups (high arousal emotions v.s. low arousal emotion). This indicates that prominence is greater for high arousal emotions than low arousal ones with statistical significance at the 0.01 level (t-statistic = 2.73, p-value = 0.0031). The difference between speakers (F-statistic: 0.02; p-value = 0.88) and speaker-emotion classes (F-statistic: 0.79; p-value = 0.53) were not statistically significant.

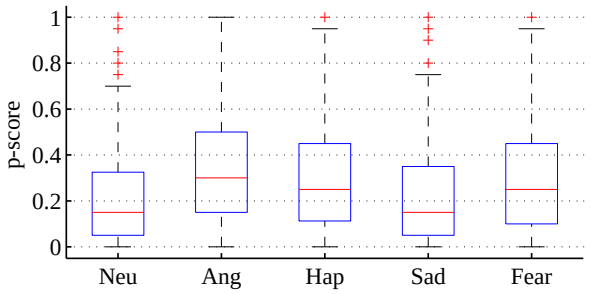

Figure 1: Boxplots of p-score (prominence score) for each emotion group

Sentence	RMS	CA	jaw	SM	f0	d_ac	d_ar
1 (spk1)	0.49	0.52	0.51	0.50	0.30	0.11	0.46
1 (spk2)	0.64	0.72	0.66	0.57	0.31	0.59	0.70
7 (spk1)	0.56	0.46	0.43	0.34	0.32	0.45	0.34
7 (spk2)	0.57	0.62	0.54	0.54	0.36	0.19	0.10

Table 1: Pearson's correlation coefficient between p-score (prominence score) and the best feature in each feature groups. 'RMS' denotes the RMS energy feature group; 'CA' denotes the critical articulator feature group; 'jaw' denotes the jaw feature group; 'SM' denotes the syllable magnitude; 'f0' denotes the f0 feature group; 'd_ac' and 'd_ar' denote the acoustic and articulatory durations respectively; 'spk1' and 'spk2' denotes speaker 1 and speaker 2, respectively. Underlined values are statistically *insignificant* at the 0.01 level.

3.2. Correlation Analysis

Next, we analyzed correlations between prominence and individual feature group. We limit our experiments to within-speaker analysis here. We first investigate linear relationships for each speaker and each sentence.

Table 1 lists Pearson's correlation coefficients between p-value and the most correlated feature in each feature group. For example, the cell corresponding to column 'RMS' and row 'Sentence 1 (spk1)' refers to the Pearson's correlation coefficient between the p-score and the most correlated feature among RMS statistics in the data of sentence 1 and speaker 1. Results indicate that the Pearson's correlation between p-score and the best feature in each group is statistically significant at the 0.01 level against the null hypothesis that there are no correlations with prominence. The table shows multiple patterns depending on sentence and feature group. First, syllable magnitude is more correlated with prominence in sentence 1 than sentence 2. It is worth noting that sentence 2 has more variations of vowels than sentence 1 (See Section 2.1), as a result of which the jaw excursion is influenced by the vowel quality (as well as syllable-level prominence) more in sentence 2 than in sentence 1. This result provides supporting evidence that vowel quality influences syllable magnitude computation significantly, which is expected. However, the amount of influence seems to vary depending on speakers; The difference is only 0.03 for speaker 2, while it is 0.16 for speaker 1. This also suggests the necessity of vowel normalization procedure for more consistent representation of syllable magnitude derived from Jaw excursion. Second, the best statistic of critical articulator is more correlated with prominence than syllable magnitude, which is a novel finding. In all four cases, the most correlated statistic was the range of critical articulator. The computation of syllable strength (i.e., prominence) in the C/D model has conventionally relied only on jaw excursion parameter. In fact, considering the best statistic feature, critical articulator shows greater (at least slightly) correlation than the jaw. This result suggests that the kinematics of the critical articulator (for the onset of the syllable) are strong indicators for representing the prominence of the corresponding syllable. Finally, the best RMS energy statistic is more correlated with prominence than the best f0 statistic in all cases. In fact, all of the RMS energy statistics had higher correlations with prominence than the best f0 statistic. On the other hand, the best RMS energy statistic seems to be comparable to the best critical articulator statistic and the best jaw statistic. Duration features, in general, seemed to have low correlations with prominence compared to other features.

Speaker	RMS	CA	jaw	SM	f0	d_ac	d_ar
1 (n)	0.52	0.47	0.45	0.40	0.25	0.30	0.38
1 (u)	0.49	0.41	0.37	0.29	0.26	0.32	0.35
2 (n)	0.58	0.66	0.59	0.55	0.28	0.38	0.39
2 (u)	0.56	0.63	0.56	0.51	0.28	0.40	0.37

Table 2: Pearson's correlation coefficient between p-score (prominence score) and the best feature in each feature groups. See the caption of Table 1 for the description of the notations for each feature group. 'n': normalized features; 'u': un-normalized features.

Next, we examined the correlations for individual feature groups after merging the two sentences' data along speakers and emotions. Table 2 shows Pearson's correlation coefficients between p-score and individual features for each speaker. We examined the same feature groups as Table 1 in this analysis. To analyze the effect of sentence level variation we report both normalized and un-normalized feature values. The normalization was performed by standardizing features (rescale to zero mean and one standard deviation) within the two sentences. As before, all values are significant against the null hypothesis that there is no correlation. The normalization of sentence level variation seems to result in higher correlations across all features, rendering credence to the process. The correlation coefficients for both sentences combined are mostly comparable to the results for individual sentences in Table 1, except for f0 and durations. This suggests that for the given data, the correlations tends to be consistent across the two sentences. When we combined data of the two sentences, f0 group tends to be significantly less correlated with prominence than the other individual articulatory-acoustic feature groups. Also, in terms of the best feature in each group, RMS, CA and jaw groups always show greater correlation coefficients than syllable magnitude, which is consistently observed in both speakers' data. However, the order of the feature group in terms of the correlations of the best feature varies depending on speakers, which may indicate that the two speakers emphasized different prosodic-articulatory controls when delivering relative strength of syllables during speech production of emotional speech.

Finally, we examined correlations for different feature groups depending on emotions. Table 3 lists Pearson's correlation coefficients between prominence for each of the five emotions and individual feature groups for each speaker. The most correlated feature group with prominence clearly varies depending on emotion but in most cases, the best RMS energy feature is highly correlated with prominence. F0 continued to

Emotion	RMS	CA	jaw	SM	f0	d_ac	d_ar
Neu (spk1)	0.59	0.44	0.39	0.26	0.26	0.24	0.29
Neu (spk2)	0.48	0.47	0.37	0.37	0.22	0.26	0.19
Ang (spk1)	0.57	0.33	0.28	0.16	0.35	0.12	0.32
Ang (spk2)	0.77	0.75	0.65	0.65	0.27	0.44	0.40
Hap (spk1)	0.60	0.42	0.35	0.32	0.27	0.32	0.37
Hap (spk2)	0.58	0.71	0.69	0.62	0.40	0.44	0.44
Sad (spk1)	0.57	0.54	0.49	0.44	0.29	0.58	0.46
Sad (spk2)	0.58	0.67	0.61	0.52	0.14	0.52	0.47
Fear (spk1)	0.59	0.42	0.46	0.43	0.22	0.45	0.47
Fear (spk2)	0.45	0.53	0.48	0.40	0.41	0.42	0.39

Table 3: Pearson's correlation coefficients between p-score and the best feature in individual feature groups, per emotion and speaker. Neu-Neutral, Aan-Angry, Hap-Happy. See the caption of Table 1 for the description of the notations for each feature group. Underlined values are statistically *insignificant* at the 0.01 level.

show low correlations with prominence. The correlations of the best f0 feature are not statistically significant at the 0.01 level in many cases (neutral, happy and fear for speaker 1; neutral, angry and sad for speaker 2). This suggests that the two speakers tend to put more weights on the controls of the other prosodic-articulatory cues than f0.

4. Conclusions and future work

This paper reports our preliminary investigation on relations between syllable-level prominence and articulatory-prosodic cues in emotional speech. Our analysis was performed in the context of the C/D model. Results indicate that prominence is highly correlated with the kinematics of linguistically critical articulator for the syllable onset. Also, the movements of critical articulator are often more correlated than the traditional measure of the syllable magnitude in the C/D model.

While the results from this initial study point to the potential benefits of incorporating other articulatory and prosodic cues for prominence representation compared to the jaw excursion measure, more detailed experiments are required to further validate the observations made in this paper. First, data of more sentence prompts and more speakers is needed to generalize our findings, including the relations between prominence and the articulatory-prosodic cues, influence of vowel quality to the syllable magnitude computation, and speaker-normalization process. Second, a larger dataset would allow us to test more sophisticated and complicated (nonlinear) models for prominence representation. In the C/D model framework, the examined articulatory-prosodic parameters are indirectly related to syllable magnitude (i.e., prominence), but the relations have not been thoroughly explored and mathematically established for the entire context of English sentences. Aspects of the C/D model that need to be explored also include the implementation of vowel normalization algorithm (e.g., [20]), and the determination of syllable nuclei when consonant onset or offset is missing. Data-driven statistical modeling is of our interest due to the complex spatio-temporal relations between these parameters and prominence. Our future work will include these directions.

5. Acknowledgements

This work was supported by NSF IIS-1116076.

6. References

[1] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *The Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991.

[2] S. J. Eady and W. E. Cooper, "Speech intonation and focus location in matched statements and questions," *The Journal of the Acoustical Society of America*, vol. 80, no. 2, pp. 402–415, 1986.

[3] A. Fernald and C. Mazzie, "Prosody and focus in speech to infants and adults." *Developmental psychology*, vol. 27, no. 2, p. 209, 1991.

[4] L. Hiyakumoto, S. Prevost, and J. Cassell, "Semantic and discourse information for text-to-speech intonation," *Concept to Speech Generation Systems*, pp. 47–56, 1997.

[5] Y. Mo, J. Cole, and E.-K. Lee, "Naïve listeners' prominence and boundary perception," in *International Conference on Speech Prosody*, Campinas, Brazil, 2008, pp. 735–738.

[6] O. Fujimura, "The c/d model and prosodic control of articulatory behavior," *Phonetica*, vol. 57, no. 2-4, pp. 128–138, 2000.

[7] P. Bonaventura, "Invariant patterns in articulatory movements," Ph.D. dissertation, Ohio State University, 2003.

[8] J. Kim, D. Erickson, S. Lee, and S. Narayanan, "A study of invariant properties and variation patterns in the Converter/Distributor model for emotional speech," in *Proceedings of Interspeech*, 2014, pp. 413–417.

[9] D. Erickson, J. Kim, S. Kawahara, I. Wilson, C. Menezes, A. Suemitsu, and J. Moore, "Bridging articulation and perception: The c/d model and contrastive emphasis," in *International Congress Phonetic Sciences*, 2015.

[10] M. E. Beckman and J. Edwards, "Articulatory evidence for differentiating stress categories," *Laboratory phonology III: Phonological structure and phonetic form*, pp. 7–33, 1994.

[11] D. Erickson, "An articulatory account of rhythm, prominence, and phrasal organization," in *Proceedings of Speech Prosody*, 2006.

[12] P. Wagner, "Evaluating metrical phonology – a computational-empirical approach," *Proceedings of Konvens—Sprachkommunikation*, vol. 161, 2000.

[13] A. Windmann, I. Jauk, F. Tamburini, and P. Wagner, "Prominence-based prosody prediction for unit selection speech synthesis," ser. Proceedings of Interspeech, Florence, Italy, 2011.

[14] G. Fant and A. Kruckenberg, "Preliminaries to the study of swedish prose reading and reading style," *STL-QPSR*, vol. 2, no. 1989, pp. 1–83, 1989.

[15] J. R. de Pijper and A. A. Sanderman, "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2037–2047, 1994.

[16] B. M. Streefkerk, L. C. Pols, and L. ten Bosch, "Acoustical features as predictors for prominence in read aloud dutch sentences used in ann's," in *EUROSPEECH*, 1999, pp. 551 – 554.

[17] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, pp. 425–452, 2010.

[18] Y. Mo, "Prosody production and perception with conversational speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2011.

[19] A. Katsamanis, M. Black, P. G. Georgiou, L. Goldstein, and S. S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA, Jan 2011.

[20] J. Williams, D. Erickson, Y. Ozaki, A. Suemitsu, N. Minematsu, and O. Fujimura, "Neutralizing differences in jaw displacement for english vowels," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3607–3607, 2013.