A study of invariant properties and variation patterns in the Converter/Distributor model for emotional speech

Jangwon Kim[1], *Donna Erickson*[2], *Sungbok Lee*[1], *Shrikanth S. Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA
[2]Kanazawa Medical University, Kanazawa, Japan

jangwon@usc.edu, ericksondonna2000@gmail.com, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

Invariant properties of vocal organ controls at an abstract level are crucial for better understanding and modeling of the speech production mechanism. Despite the large variability of articulatory movements at the execution level, the Converter/Distributor (C/D) model provides a systematic and comprehensive framework for the prosodic organization of speech production, based on the invariant properties of articulatory movements with the concept of "iceberg" region. The goal of this paper is two-fold: (i) to examine the invariant properties in the C/D model in emotional speech, and (ii) to understand emotion-dependent variation patterns of important parameters in the C/D model framework. Experimental results support the validity of strong linear relationship between the speed and excursion of critical articulators at the iceberg points for emotional speech. Also, emotion-dependent variation patterns of the C/D model parameters, (e.g., relatively smaller "shadow" angle and greater syllable magnitude for happiness) are reported. Finally, the emotion-dependent relationships between the *abstract*-level C/D model parameters and the *surface*-level parameters of the invariant articulatory behaviors are reported.

Index Terms: emotional variation, temporal organization of speech, C/D model, invariant property

1. Introduction

The human speech signal is produced by the coordinated controls of vocal organs [1, 2] with large variability on their surface movements [3, 4]. One of the main challenges in speech production modeling is, therefore, to represent articulatory behaviors in an effective, but simpler way. Despite the large variability of articulatory movements, previous studies have reported the presence of relatively invariant portions, called "iceberg" regions [5, 6], of the transient articulatory trajectories of demisyllables. More specifically, it has been observed that the speed of the (linguistically) critical articulator for producing the consonant in the demisyllable is relatively invariant at a certain excursion point regardless of prosodic change, e.g., different level of stress on the syllable, as long as the vowel of the demisyllable and para-linguistic factors, e.g., speaker-specific characteristics, gender and emotion, are fixed. The iceberg region is roughly the fastest part of the critical articulatory trajectory in each demisyllable. This invariant characteristic at the iceberg is considered in the Converter/Distributor (C/D) model.

The C/D model is a comprehensive model of the speech production system, a part of which describes the (abstract) high-level temporal organization of speech, based on articulatory movements [7]. In this model, sequential syllable pulses represent the rhythmic pattern of consecutive syllables in the utterance. Syllable triangles are constructed based on the syllable pulses, where the height of the triangle reflects the syllable magnitude, i.e., syllable prominence, and the length of the base of the triangle reflects *abstract* syllable duration in the articulatory domain. Para-linguistic factors, e.g., speaker style, rate of speech and emotion, affect the variation of syllable pulse trains, resulting in the variation of the amplification and timing of the impulse response function (IRF) for consonantal gestures [8, 9]. The IRFs are prototype time functions that represent inherent characteristics of elemental consonantal gestures, e.g., apical stop, labial fricative, velar stop. See [6, 9, 10] for the details of the entire C/D model that contains other components needed for generating articulatory signals from these variables.

The present study investigates the invariant properties and variation patterns of articulatory movements of emotional speech in the perspective of the C/D model. Such knowledge is valuable from both a theoretic standpoint (to shed further light on the articulatory control mechanism with emotion coloring) and application perspectives (such as in informing better articulatory modeling and (re-)synthesis with emotion). The invariant properties in the C/D model include (i) the strong linear relationship between the (vertical) excursion of critical articulators and the articulatory speed at the iceberg and (ii) the linear relationship between the syllable duration and the syllable magnitude. The latter is based on the assumption that the acute angles of the two side lines of all syllable triangles, called "shadow" angles, are identical. This we refer to here as the consistency assumption.

The present study examines the variation of the timing and amplitude of the iceberg points found in the articulatory trajectories and the "shadow" angle. In the C/D model framework, emotion affects the parameters of IRFs, not the surface articulatory trajectories directly. However, the IRFs represent abstract articulatory gestural controls that are not directly observable due to the highly nonlinear nature between the IRFs and articulatory signals [11], which makes the direct analysis on the IRFs harder. This study examines the variation of articulatory parameters that are influenced by the change of the IRFs as a function of emotion.

2. Methods

2.1. Datasets

The ElectroMagnetic Articulography (EMA) dataset collected by the NDI WAVE system is used in this study. A sentence "*Pam said bat that fat cat at that mat*" was spoken by a female native speaker of American English. The stimulus was designed specifically for the study of the C/D model. For consonants, it contains only stops and fricatives in which the invariant prop-

Table 1: *Confusion between the target emotion and the final emotion label, i.e., the best (perceived) emotion. 'Neu' is neutrality, 'Ang' is anger, 'Hap' is happiness, 'Sad' is sadness.*

		Final label					
		Neu	Ang	Hap	Sad	Fear	Other
Target	Neu	5	0	0	0	0	0
	Ang	0	5	0	0	0	0
	Hap	0	0	5	0	0	0
	Sad	0	0	0	5	0	0
	Fear	0	0	0	1	4	0
	Total	5	5	5	6	4	0

erties of the C/D model have been shown in literature, e.g., in [7, 9]. For vowels, it has only two vowels, eight /AE/ and one /EH/, so that the variation of the C/D model parameters due to vowels is minimized. The sentence was repeated five times for each of the five emotions, such as neutrality, anger, happiness, sadness and fear. The speaker was a professional actress who had theatrical vocal training. She was asked to start speaking after she had immersed herself in the target emotion.

A 6 degree-of-freedom (DOF) sensor of the NDI WAVE system was used as the reference sensor, and six 5-DOF sensors were used for monitoring the movements of articulators, such as the tongue tip (TT), the tongue blade (TB), the tongue dorsum (TD), the upper lip (LL), the lower lip (UL) and the jaw. The 3-dimensional coordinates of the six 5-DOF sensors were recorded at a sampling rate of 100 Hz, and speech waveform was simultaneously recorded at a sampling rate of 22050 Hz. Occlusal plane correction was performed on the articulatory data of all utterances by using the recording of three 5-DOF sensors attached on the bite plate. After interpolating missing frames by the piecewise cubic Hermite interpolating polynomial, each sensor trajectory was smoothed with a 9th-order Butterworth low pass filter with a cutoff frequency of 20 Hz. Only tongue tip, tongue dorsum and lower lip sensors were selected as critical articulatory sensors for the sake of simplicity of analysis along with jaw contribution.

The best emotion of each utterance was judged by 11 native speakers of American English. After listening to each utterance, the evaluators were asked to choose (1) the best representative emotion among six categories, such as neutrality, anger, happiness, sadness and 'other,' where 'other' was for the case that none of the listed five emotions was the best, (2) confidence in their judgment, and (3) the strength of emotion expression. Confidence and strength were evaluated on a five-point Likert scale. The best emotion was determined by majority voting. If there were multiple emotions with the same evaluation score, the one of higher mean of confidence scores was chosen. Table 1 shows the confusion between the target emotion and the best (perceived) emotion used for analysis.

2.2. Parameter extraction

The critical articulators should be defined for computing C/D model parameters pertinent to this study. In this study, the critical articulator for each phone is determined based on the place of articulation, i.e., the tongue tip for coronals (/S/,/TH/,/T/,/D/), the lower lip for labials (/P/,/M/,/B/,/F/), and the tongue dorsum for dorsals (/K/). Although there is no initial consonant for "AT," the final consonant /T/ of the previous word "CAT" was used for extracting C/D model parameters, because "CAT" and "AT" were spoken continuously without pause.

In most literature (e,g., [6, 10, 12]), the iceberg point is algorithmically determined at the minimum variance point of a

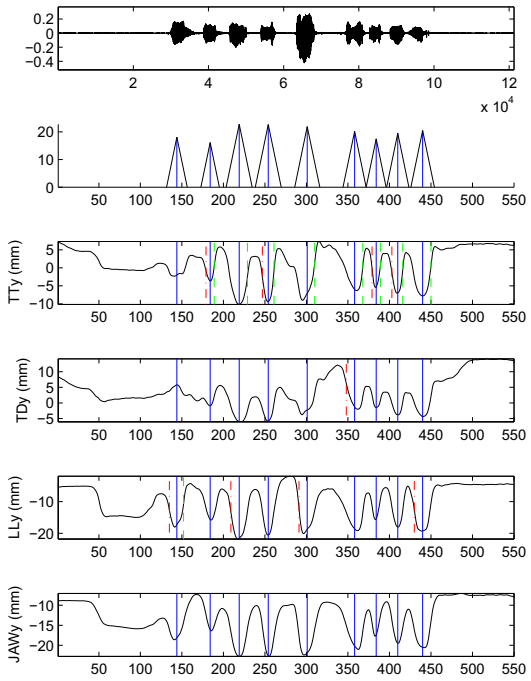

Figure 1: Syllable triangles constructed for a neutral utterance of "Pam said bat that fat cat at that mat." The 1st panel is the speech waveform. The 2nd panel shows syllable triangles. In the other panels, the red dash-dot line denotes the iceberg time point for onset; the green dashed line denotes the iceberg time point for coda; the blue solid line denotes the syllable center point.

number of trajectories of the same demisyllable. One approach is to find the point of the minimum root-mean-squared-error in the horizontal direction after optimal time shifting of the trajectories to the reference trajectory [5, 8]. Another approach is to choose the point of the minimum "iceberg metric" among multiple vertical movement bands of the critical articulator [10]. The iceberg metric is proportional to the variance of articulatory speed and inversely proportional to the mean of articulatory speed in the band. Although these algorithmic approaches can find reliable iceberg points abiding in the invariability principle of the C/D model, these methods require a large number of trajectory samples to secure the reliability. However, the number of trajectories of each demisyllable and each emotion is very limited in the present study. In the present study, therefore, the iceberg point is, therefore, determined at the maximum speed point of the critical articulator for the onset or coda of each demisyllable as in [13].

The midpoint between the two iceberg points (for onset and coda) in each syllable is where the syllable pulse was placed. The excursion of the jaw at the midpoint was considered to be the height of the syllable pulse, which represents the syllable magnitude. The excursion of an articulator refers to the shortest distance between the occlusal plane and the position of the articulator [6]. Then, the "shadow" angle of the triangle was calculated for each utterance in such a way that there is at least one pair of the close edges of adjacent triangles which meet with no overlap in between any adjacent triangles [7]. Figure 1 illustrates the iceberg time points, and syllable centers, syllable triangles in a neutral speech utterance. The time difference between the onset/coda pulse, i.e. syllable triangle edge, to the iceberg point of the demisyllable is refered to as τ (not shown in the figure, but discussed in Sec. 4).

Table 2: *'**' denotes that p-value < 0.0000005. '*' denotes that p-value < 0.00005. N=25.*

Syllable	CV				VC			
	β_1	β_2	F	p	β_1	β_2	F	p
PAM	12.1	-5.1	405	**	9.8	5.0	92	**
SAID	11.4	2.6	78	**	7.8	57.7	68	**
BAT	6.8	70.8	98	**	16.0	-83.2	92	**
THAT	15.0	-35.0	187	**	13.3	1.4	64	**
FAT	10.1	25.6	238	**	10.9	10.7	53	**
CAT	12.7	-52.6	48	**	12.2	12.8	124	**
(T) AT	11.9	23.3	139	**	10.6	43.8	26	*
THAT	14.9	-26.7	57	**	13.5	-1.3	88	**
MAT	11.1	-2.0	285	**	16.5	-66.2	55	**

3. Analysis on the invariant properties of the C/D model

This section discusses two invariant properties in the C/D model associated with emotional speech. One is the strong linear relationship between the excursion of the critical articulator and the speed of the articulator at the iceberg, and the other is the consistency of the "shadow" angle values across utterances spoken with the same emotion. These invariant properties are examined across emotions and within emotion.

3.1. Invariant properties at iceberg points

Visual inspection of the scatter plot (Figure 2) shows a strong linear relationship between the excursion of the critical articulator and the speed of the articulator at the iceberg point for CV and VC demisyllables, for all utterances regardless of the emotion condition. For demisyllables, happiness shows the greatest excursion and the highest articulatory speed, while neutrality shows the smallest excursion and the lowest articulatory speed. A linear regression analysis (Table 2) shows the F-statistic and p-value for all emotion conditions, including neutrality, for each CV/VC demisyllable. In Table 2, the p-value is significant at α = 0.00005 level in all cases, indicating that a linear relationship between the two parameters is maintained across all emotion conditions, not just in neutral speech. This support of the C/D Model assumption of the linearity of articulatory speed and excursion is discussed further in Sec. 5.

3.2. Invariance of the shadow angle of the syllable triangle

Next, we examine the shadow angle within each emotion condition in order to investigate the invariance of the shadow angle of the syllable triangle.

Figure 3 shows the errorbar plot of the shadow angle computed for each utterance. Note that the angle varies depending on the emotion: 36 degrees for neutrality, 32 for anger, 24 for happiness, 30 for sadness, and 30 for fear. The standard deviation of the angle is smallest for neutrality (0.93), and significantly greater for the other emotions: 3.31 for anger, 2.67 for happiness, 2.89 for sadness, and 3.48 for fear. This suggests that for a neutral speech condition, the shadow angle is fairly consistent, while it is relatively variable for emotional speech, within emotion as well as across emotions. This is an interesting finding and will be discussed in more detail in Sec. 5.

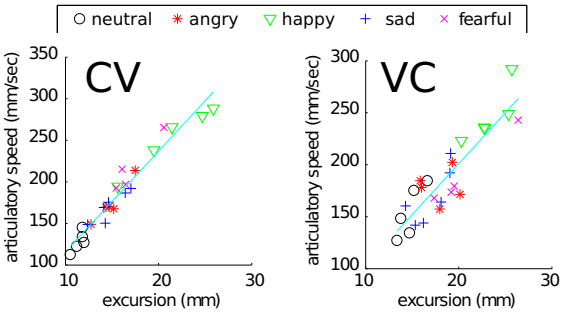

Figure 2: Example scatter plots for the excursion of the critical articulator (of consonant) and the articulatory speed at icebergs in CV/VC demisyllables. "*Pam*" is used in this plot.

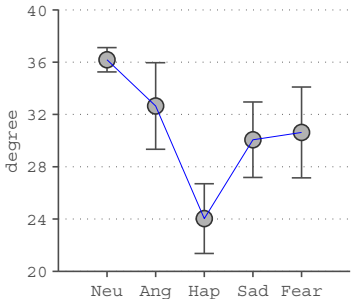

Figure 3: Errorbar plot of the "shadow" angle for each emotion.

4. Analysis of emotional variability in the C/D model

In the C/D Model framework, emotional variation factors are a part of the utterance parameters which cause variation of syllable magnitudes and IRF parameters (i.e., phase and magnitude of IRF peak). The variation of IRFs parameters, such as amplification (affected by the syllable magnitude) and timing (from the onset/coda excitation pulses), affect consonantal gestures. It follows from this that (i) the time-shifting of the IRFs influences the location of the maximum speed time points of the critical articulators and (ii) the amplification of the IRFs influences the speed (i.e., increases the speed) of the critical articulators at the iceberg point. Since the IRFs are hidden, the present paper analyzes the surface phenomenon directly. The goal is to understand the effects of emotion to the relative timing and speed of the iceberg points in syllables.

First, we investigate the effects of emotion on syllable magnitudes. Figure 4 shows the syllable magnitudes for each syllable, for each emotion condition. Overall, happiness shows the greatest syllable magnitude (jaw displacement), while anger shows the smallest. It would seem for happiness, the speaker uses greater jaw movement and for anger, this speaker speaks with a "clenched jaw," a term often used in novels to describe expressions of cold anger. Note that although the syllable magnitude is smaller for anger compared to that of the other emo-

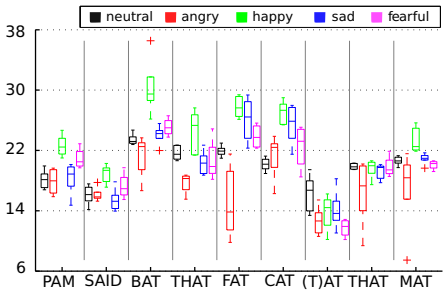

Figure 4: Syllable magnitude, as jaw excursion, for each monosyllabic word in the utterance

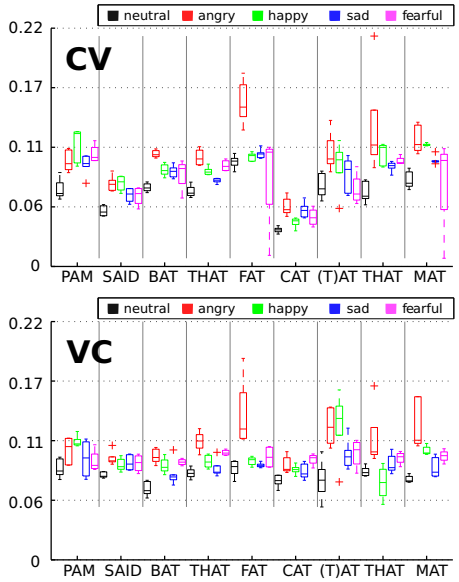

Figure 5: Ratio of articulatory speed (at the iceberg point for CV/VC demisyllable) to the syllable magnitude for each demisyllable.

tions, the speed and excursion of the critical articulators (as shown in Figure 2) is not significantly smaller. This finding hints that the emotional factor, e.g., the one resulting in the clenched jaw for anger, causes the variation of the relationship between the syllable magnitude and the amplitude of articulatory gesture (exhibited in the speed of the critical articulators at the iceberg point).

We further investigated the relationship between the speed of critical articulator and the syllable magnitude for different emotions. Figure 5 shows the ratio of the speed of the critical articulator (at CV/VC iceberg point) to syllable magnitude for each sample. This figure indicates that the ratio varies significantly depending on emotion. Note that the syllable magnitude is an indicator of syllable prominence in the articulatory domain. Also, note that the articulatory speed at the iceberg point is the maximum speed value of the critical articulator. Overall, the ratio for emotional speech (anger, happiness, sadness, fear) is greater than the ratio for neutral speech, indicating that the ratio of the releasing speed of the critical articulator to the syllable magnitude is greater when the subject is emotional. This implies that the speaker tends to articulate with stronger consonantal gestures for critical articulators when the person is emotionally charged. This tendency is more consistent across CV demisyllables than VC demisyllables. In sum, results suggest that the maximum speed of critical articulators given syllable prominence varies depending on emotion.

Finally, we examined the time variation (τ) between the onset pulse, i.e., syllable triangle edge, to the iceberg point of CV demisyllable. Note that τ should be the same as the time difference between the coda and the iceberg point of VC demisyllable. This information is useful in the sense that it directly relates the abstract representation for temporal structure of an utterance to the surface phenomenon of articulatory movements. Figure 6 shows box plots of τ for each syllable. The mean of τ is greater for neutrality than for the other emotions in all cases, except 'CAT.' τ is a function of the shadow angle and the syllable magnitude: A larger shadow angle and greater syllable magnitude cause greater τ, which is in line with our previous observations in Figure 3 and Figure 4. For example, anger shows a smaller shadow angle and smaller syllable magnitude (due to idiosyncratic "clenched jaw" of the speaker) than neutrality, so τ of anger is also smaller than τ of neutrality.

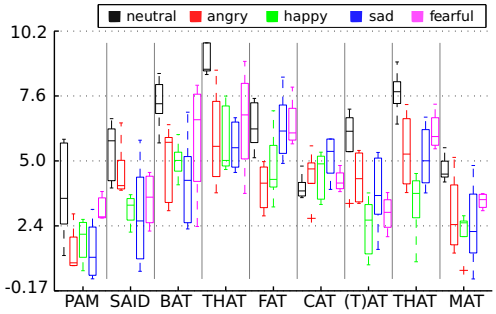

Figure 6: The top panel shows the time difference between the onset pulse point and the iceberg point.

5. Discussion and Future work

In the analyses of this paper, we observed that emotion influences the shadow angle, syllable magnitude, the ratio of the maximum speed of the critical articulator in demisyllables to the syllable magnitude, and τ. One hypothetical reason for the variation of the shadow angle is that the assumption of the linear dependency between the syllable magnitude and the articulatory syllable duration is not valid in emotional speech. More specifically, jaw excursion may not be linearly dependent on the syllable duration in emotional speech, e.g., in clenched jaw for anger. This may point to the need of more comprehensive representation for the syllable magnitude in the C/D model framework for emotional speech. Another hypothetical reason is the conventionally applied assumption that the angles of the CV demisyllable and the VC demisyllable are identical is not valid in emotional speech. A previous study has raised the possibility of the two angles' asymmetry in the phrase-final elongation [6]. In fact, the symmetry has been assumed for the simplicity of analysis, not for the theoretic or algorithmic neccesity in the C/D model framework.

According to the C/D model, smaller shadow angle given the same syllable duration indicates greater syllable magnitude, thereby greater jaw excursion and faster maximum speed of critical articulator. This finding is in line with the faster articulatory movement and the greater movement range for happiness reported in previous studies [14, 15, 16, 17].

It should be noted that the variation of these parameters mentioned above is not independent from each other. The shadow angle is a function of the syllable magnitude and the time gap between the closest edges of adjacent syllable triangles. The maximum speed of the critical articulator is a function of the IRFs, which is affected by the syllable magnitude. τ is also a function of the IRFs. Hence, a *joint* analysis and modeling for the variation of these parameters as a function of emotion, is important to represent emotional variability in the C/D model framework. An articulatory re-synthesis experiment with emotion transformation can be useful for evaluating the joint model. These constitute future work to be explored.

6. Acknowledgements

This work was supported by NSF IIS-1116076, NIH DC007124, and the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (C)#22520412 and (C)#25370444. Special thanks to Mary Francis for her devotion and help in all SAIL research efforts.

7. References

[1] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[2] C. A. Fowler and E. Saltzman, "Coordination and coarticulation in speech production," *Language and Speech*, vol. 36(2,3), pp. 171 – 195, 1993.

[3] L. L. Koenig, J. C. Lucero, and E. Perlman, "Speech production variability in fricatives of children and adults: Results of functional data analysis," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3158–3170, 2008.

[4] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695–710, 2009.

[5] O. Fujimura, *Relative invariance of articulatory movements: An iceberg model*. Lawrence Erlbaum Assoc., Hillsdale, NJ, 1986, ch. 11, pp. 226–242, in Invariance and Variability of Speech Processing, edited by J. S. Perkell and D. Klatt.

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

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

[8] ——, "Syllable timing computation in the C/D model," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan, September 1994, pp. 519 – 522.

[9] ——, "Temporal organization of speech utterance: A C/D model perspective," *Cadernos de Estudos Lingüísticos*, vol. 43, pp. 9 – 35, 2002.

[10] C. Menezes, "Rhythmic pattern of American English: An articulatory and acoustic study," Ph.D. dissertation, Ohio State University, 2003.

[11] O. Fujimura, "C/D model: A computational model of phonetic implementation," *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 17, pp. 1–20, 1994.

[12] P. Bonaventura and O. Fujimura, *Articulatory movements and phrase boundaries*, ser. Oxford linguistics. OUP Oxford, 2007, ch. 14, pp. 209–227, in Experimental Approaches to Phonology, edited by Sole, M.J. and Beddor, P.S. and Ohala, M.

[13] D. Erickson, "More about jaw, rhythm and metrical structure," in *Acoustical Society of Japan, fall meeting*, 2010, p. 103.

[14] J. Kim, S. Lee, and S. S. Narayanan, "A study of interplay between articulatory movement and prosodic characteristics in emotional speech production," in *Proceedings of Interspeech*. ISCA, pp. 1173–1176.

[15] S. Lee, S. Yildirim, A. Kazemzadeh, and S. S. Narayanan, "An articulatory study of emotional speech production," *Proceedings of Interspeech*, pp. 497 – 500, 2005.

[16] J. Kim, S. Lee, and S. S. Narayanan, "An exploratory study of the relations between perceived emotion strength and articulatory kinematics," in *Proceedings of Interspeech*, 2011, pp. 2961 – 2964.

[17] S. Lee, E. Bresch, J. Adams, A. Kazemzadeh, and S. S. Narayanan, "A study of emotional speech articulation using a fast magnetic resonance imaging technique," in *Proceedings of Interspeech*, Pittsburgh, PA, September 2006, pp. 2234 – 2237.