A kinematic study of critical and non-critical articulators in emotional speech production

Jangwon Kim,[a] Asterios Toutios, Sungbok Lee, and Shrikanth S. Narayanan
Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089

(Received 11 November 2013; revised 10 December 2014; accepted 5 February 2015)

This study explores one aspect of the articulatory mechanism that underlies emotional speech production, namely, the behavior of linguistically critical and non-critical articulators in the encoding of emotional information. The hypothesis is that the possible larger kinematic variability in the behavior of non-critical articulators enables revealing underlying emotional expression goal more explicitly than that of the critical articulators; the critical articulators are strictly controlled in service of achieving linguistic goals and exhibit smaller kinematic variability. This hypothesis is examined by kinematic analysis of the movements of critical and non-critical speech articulators gathered using eletromagnetic articulography during spoken expressions of five categorical emotions. Analysis results at the level of consonant-vowel-consonant segments reveal that critical articulators for the consonants show more (less) peripheral articulations during production of the *consonant-vowel-consonant* syllables for high (low) arousal emotions, while non-critical articulators show less sensitive emotional variation of articulatory position to the linguistic gestures. Analysis results at the individual phonetic targets show that overall, between- and within-emotion variability in articulatory positions is larger for non-critical cases than for critical cases. Finally, the results of simulation experiments suggest that the postural variation of non-critical articulators depending on emotion is significantly associated with the controls of critical articulators.
 [http://dx.doi.org/10.1121/1.4908284]

[CYE] Pages: 1411–1429

I. INTRODUCTION

Previous studies have shown that the emotional states of speakers influence the acoustic and articulatory characteristics of their speech. While studies on the acoustic properties of voice quality and prosody of emotional speech abound in the literature (Gobl and Ni'Chasaide, 2003; Ververidis and Kotropoulos, 2006; Williams and Stevens, 1972; Lee and Narayanan, 2009; Yildirim *et al.*, 2004; Busso *et al.*, 2009), there are considerably fewer studies about articulatory details of emotional speech, presumably due to difficulties in obtaining direct articulatory data. Although various techniques, such as x-ray microbeam (Fujimura *et al.*, 1973), ultrasound (Stone, 2005), electromagnetic articulography or EMA (Iskarous *et al.*, 2010; Perkell *et al.*, 1992), and magnetic resonance imaging or MRI (Narayanan *et al.*, 1995; Narayanan *et al.*, 2004) have been used for the study of speech production, the data collection environment is not ideal for investigating natural emotion expression in speech. Nevertheless it has been shown that distinctive emotional information is present in EMA data of *elicited* and *acted* emotional speech (Erickson *et al.*, 2004; Erickson *et al.*, 2006; Lee *et al.*, 2005). In particular, it has been reported that the position and speed of articulators, particularly their properties at the syllable, word, and utterance levels, are important emotional features (Lee *et al.*, 2005; Lee *et al.*, 2006; Kim *et al.*, 2009, 2011; Kim *et al.*, 2012). Preliminary results on the interplay between prosodic characteristics and articulatory movements as a function of emotion have also been reported (Kim *et al.*, 2010).

Despite progress in understanding the articulatory aspects of emotional speech, there is still limited knowledge about the behavior of individual articulators for achieving emotional goals in parallel with linguistic goals during speech production. The present study aims at investigating the relationship between the variability in the kinematic behavior of specific articulators for the achievement of specific linguistic and emotional goals during speech production. A better understanding of such details can shed further light on intra-speaker variability in speech production. Such knowledge can also benefit modeling and synthesis of emotional speech.

Linguistic criticality of articulators is an important factor to characterize articulatory variability during speech production. For achieving certain linguistic goals, (linguistically) critical articulators are more carefully controlled and display less variability, than non-critical articulators. For example, in producing /t/, it is essential that the tongue tip comes in instantaneous contact with the alveolar ridge, while the positions of the lips and the tongue dorsum may be more variable. In this case, the tongue tip is considered the critical articulator for /t/, while the lips and the tongue dorsum are considered non-critical articulators. The linguistic criticality of articulators can also be categorized based on the direction of movements, which depend on constriction locations. For example, horizontal constriction of the tongue body is critical for pharyngeal vowels, while vertical constriction is critical for palatal vowels.

[a]Author to whom correspondence should be addressed. Electronic mail: jangwon@usc.edu

The present study investigates the roles of non-critical articulators for emotional information encoding. Ananthakrishnan and Engwall (2008) reported that movement information from the critical articulators alone may be enough to almost fully encode the linguistic message. They showed that from the viewpoint of linguistic encoding, the movements of non-critical articulators reflect temporal linguistic context, displaying interpolative motions of preceding and following critical points under vocal-tract physiological constraints. From the viewpoint of emotional encoding, we hypothesize that the motions of non-critical articulators may encode emotion-distinctive information in tandem with the aforementioned interpolative movements. In the present paper, we test this hypothesis on the basis of the binary distinctions of articulator criticality in realizing speech gestures within the framework of articulatory phonology (Browman and Goldstein, 1989) and implemented in the task dynamics application model (Nam *et al.*, 2004). Further details of this setup is provided in Sec. III.

We thus investigate the following three questions: (a) In what ways do the kinematics of critical and non-critical articulators vary as a function of emotions? (b) Are the kinematics of non-critical articulators more emotion-distinctive compared to the kinematics of critical articulators? (c) Is the emotion-dependent variability of non-critical articulators simply mechanical outcome of controls on critical articulators? To address the first and second questions, we analyzed articulatory variability as function of emotion at the syllable and phone levels. At the syllable level, several static and dynamic articulatory parameters extracted from EMA data at manually labeled phone-target and transition points were analyzed using distribution plots. Phone-level analysis focused on understanding the variability of task-oriented articulatory trajectory formation in emotional speech. The emotional variation in articulatory position at phone targets was quantified by the average of centroid distance and mean dispersion, which represent the variability between emotions and within each emotion, respectively. Critical and non-critical cases are compared across phones using these two standardized measures. To address the third question, we compared the emotion-dependent postural variation of true and estimated articulatory trajectories, where the estimated trajectories were generated as a function of the movements of only critical articulators. High similarity of the emotion-dependent variation of the two trajectories (true and estimated) implies considerable dependency of the emotional variation of non-critical articulators to controls of critical articulators. The articulatory model proposed in our previous study (Kim *et al.*, 2014b) was used for this simulation experiment.

In Sec. II of this paper, we describe our methods for collecting, evaluating, and processing articulatory data of emotional speech. In Sec. III, we explain our binary categorization of articulators in terms of their linguistic task criticality. In Secs. IV and V, we present our methods and results for the syllable level and phone level analyses, respectively. In Sec. VI, we present our methods and results of the simulation experiment. In Sec. VII, we provide a discussion of the results.

II. DATABASE

A. Data description

This study uses an articulatory database of emotional speech production collected at the University of Southern California. EMA was used for the data collection as described in Kim *et al.* (2011). The database includes speech waveforms, sampled at 16 kHz, and corresponding three-dimensional (3D) coordinates of six sensors attached to oral articulators, sampled at 200 Hz. Figure 1 shows the placement of the six sensors in the mid-sagittal plane. Three sensors were placed on the tongue surface: The front-most sensor (noted as *tongue tip* in the figure) was placed about 0.5–1 cm behind the anatomical tongue tip for monitoring the movement of the tongue tip as well as minimizing its interference on the articulatory action, the rear-most sensor (noted as *tongue dorsum* in the figure) was attached as far back as possible for speakers (approximately 4–4.5 cm behind the tongue tip sensor) for capturing the movements of the tongue dorsum; and the third sensor (noted as *tongue blade*) was positioned between the tongue tip and tongue dorsum sensors (about 2.5–3 cm behind the tongue tip sensor). Sensors were also glued on the upper and lower lips. Finally, a sensor was attached on the lower incisor for monitoring the movement of the jaw. The trajectories of the six sensors were recorded with a Carstens AG500 EMA system. Three native speakers of American English, one male (referred to as SB) and two female (JR and JN) produced speech in five acted categorical emotions (neutrality, hot anger, cold anger, happiness, and sadness) and three speaking styles (normal, loud, fast). The speakers had previous training in theater and acting. They were instructed to read four or five repetitions of seven sentences in every combination of the five emotions and the three speaking styles. See the Appendix for the seven sentences. There were 524 utterances for JR (7 sentences $\times$ 5 emotions $\times$ 3 styles $\times$ 5 repetitions $-$ 1 erroneous recording that was discarded), 440 utterances for JN (7 sentences $\times$ 5 emotions $\times$ 3 styles $\times$ 4 or 5 repetitions $-$ 2 erroneous recordings), and 417 for SB (7 sentences $\times$ 5 emotions $\times$ 3 styles $\times$ 4 repetitions $-$ 3 erroneous recordings). Fast style utterances were excluded from analysis because variation due to the intended speaking rate change was not within the scope of the present study. On the other

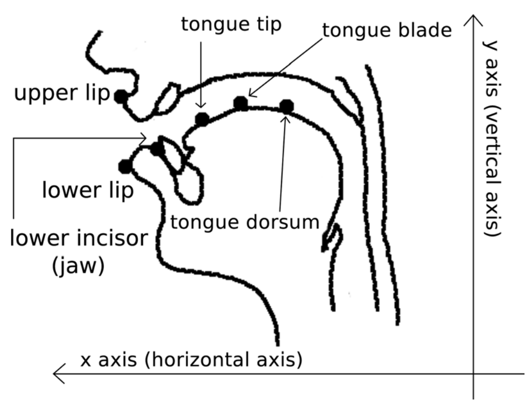

FIG. 1. Placement of EMA sensors in the mid-sagittal plane.

TABLE I. Confusion matrix between evaluated emotion determined by majority voting and target emotion of speakers. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness, and Sad is sadness. The numbers in bold are the greatest of evaluated emotion cell for each target emotion, each speaker and each intended style.

		Evaluated emotion														
		JN					JR					SB				
Target emotion	Emo	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad
Normal	Neu	**28**	0	2	0	4	**33**	0	1	0	0	**28**	0	0	0	0
	Han	0	**13**	12	4	0	1	10	**20**	0	0	4	0	**12**	1	1
	Can	0	1	**28**	0	1	0	8	**18**	0	1	0	8	**18**	0	1
	Hap	1	0	0	**33**	0	0	0	0	**23**	10	7	0	1	**20**	0
	Sad	0	0	2	0	**28**	0	0	0	0	**35**	0	0	0	0	**28**
Loud	Neu	5	8	**15**	0	0	**32**	0	1	0	0	**26**	0	1	0	1
	Han	0	**21**	2	0	0	0	**21**	9	0	0	0	**28**	0	0	0
	Can	0	6	**19**	0	0	0	9	**16**	0	0	0	10	**18**	0	0
	Hap	0	3	0	**18**	0	0	2	1	**26**	1	0	2	0	**24**	0
	Sad	0	0	1	1	**25**	0	0	0	3	**31**	0	0	0	0	**28**

hand, loud style utterances were included because loudness of speech is an important factor of emotion expression and perception especially for distinguishing emotions in the arousal dimension (Kim *et al.*, 2011).

The 3D coordinates of the six sensor position data were corrected for head movement, and the orientation of articulatory trajectories was fixed to the occlusal plane. The orientation of the occlusal plane was measured using a half-rounded bite plane on which three sensors were attached. We use the projections of the EMA sensors on the (horizontal) x axis and the (vertical) y axis shown in Fig. 1. Each raw articulatory trajectory (evolution of sensor position projected on the x or y axis) was smoothed by a ninth-order Butterworth filter with a 15 Hz cutoff frequency as in Lee *et al.* (2005).

B. Emotion evaluation

The emotion expressed in each audio utterance spoken by the three speakers of the database was evaluated by either four or five listeners, native speakers of American English and either undergraduate or graduate students at the University of Southern California (see http://sail.usc.edu/data/ema_eval_jr_short for the evaluation interface). For each speech audio, listeners were asked to choose (1) the best-representative emotion among six emotion categories (neutrality, hot anger, cold anger, happiness, sadness, and other), (2) the degree of confidence in their evaluation, and (3) the strength of emotion expression. Only speech audio was provided to the listeners in a randomized order without showing any intended goal (loudness and emotion) of the speakers. They were asked to choose "other" when none of the five given emotion categories was a good match to what they perceived. Confidence and strength were evaluated on a five-point Likert scale.

The most representative emotion for each utterance was determined by majority voting. If two emotions had the same evaluation scores, then the one with higher confidence score was chosen. The confidence of each evaluator was normalized by z-scoring across all utterances. Utterances that did not satisfy the majority voting criteria were discarded to maximize distinctive articulatory characteristics among the five categorical emotions for the present analysis. In the end, 312 utterances of JR's data, 281 utterances of JN's data, and 267 utterances of SB's data were used for analysis. It needs to be noted that the non-selected utterances are still important data for emotional speech research because they may reflect ambiguous displays and heterogeneity in judging emotions. For example, it would be important to understand what varies for emotional speech production and perception from one speaker to another because of individual differences in terms of gender, personality, and prior experience.

Table I shows the confusion matrix among emotions perceived by the judges (i.e., the result of majority voting) and emotions intended by the speakers (target emotions). The matching ratios between intended and evaluated emotions are comparable to previous studies (Grimm *et al.*, 2007; Shami and Verhelst, 2007). In addition, we observed a significant degree of confusion between hot anger and cold anger across all speakers, indicating that the two emotions are quite similar in terms of perception and/or expression. Loudly spoken (intended) neutral speech was often perceived as cold or hot anger, especially for JN's data. Also judges had a preference for hot anger to cold anger in loud style speech. These are in line with the observations of previous studies (e.g., Kim *et al.*, 2011), which noted that loudness of speech is an important factor of emotion expression and perception, especially for distinguishing categorical emotions in the arousal dimension. Finally, loudly spoken (intended) happy speech was often perceived as hot anger, presumably due to the fact that they are close in the arousal dimension of emotion perception.

III. CRITICAL AND NON-CRITICAL ARTICULATORS IN SPEECH

In the present study, the linguistic criticality of speech articulators for the realization of each phone is determined on the basis of the framework of articulatory phonology (Browman and Goldstein, 1989) and its computational ancillary task dynamics model (Saltzman and Kelso, 1987; Saltzman and Munhall, 1989). Articulatory phonology views the speech production process as composed of articulatory gestures. Specifically, the formation and release of constrictions in the vocal tract is represented by gestures depending on linguistic context, hence lexical items are differentiated by different gestural composition. The gestures are defined in terms of specific tract variables (e.g., lip aperture, tongue tip constriction degree) in task dynamics, which specifies the sets of articulators contributing to each tract variable parameters. In our subsequent analysis, we consider as critical articulators for a given phone those that the task dynamics application model (Nam *et al.*, 2004) regards as involved in the production of that phone; the rest are considered non-critical. Alternative ways of specifying the linguistic criticality of articulators have been previously proposed. Notably, Jackson and Singampalli (2008) proposed an empirical approach based on Kullback–Leibler divergence based on the assumption that the variance of articulatory positions in the mid-sagittal plane is smaller for critical articulators than non-critical articulators. The validity of this assumption has been supported by several experimental results Papcun *et al.* (1992), Frankel and King (2001), and Jackson and Singampalli (2008). This was done, however, only for the case of neutral speech; validity in para-linguistic quality, such as emotion, was never considered.

For consonants, we consider that the lower lip, tongue tip, and tongue dorsum sensors correspond to labial, apical, and dorsal articulators, respectively. Table II shows the list of consonants used for analysis and the sensors corresponding to their critical or non-critical articulators. Even though the motions of both upper and lower lips are considered important for the production of bilabial consonants, we consider here only lower lip sensor as critical for the sake of analytic simplicity, taking also into account that the motion of the upper lip is highly correlated with the motion of the lower lip during constriction and releasing gestures for the bilabial consonants.

TABLE II. List of stop and fricative consonants in the EMA database and the flesh point sensors of critical articulators of them. Note that /s/ and /z/ have two critical articulators because both tongue tip constriction and tongue dorsum wide opening gestures are critical for the production of the phones (Nam *et al.*, 2004). The list of vowels in the EMA database is [ɑ, æ, ə, ɔ, ʌ, aʊ, aɪ, ɛ, eɪ, ɪ, i, oʊ, ɔɪ, u]. The flesh-point sensors corresponding to critical or non-critical articulators of vowels is not specified here due to their less clarity than consonants.

Phone	Critical articulator	Non-critical articulator
d	Tongue tip	Tongue dorsum, lower lip
ð	Tongue tip	Tongue dorsum, lower lip
f	Lower lip	Tongue tip, tongue dorsum
g	Tongue dorsum	Tongue tip, lower lip
k	Tongue dorsum	Tongue tip, lower lip
m	Lower lip	Tongue tip, tongue dorsum
n	Tongue tip	Tongue dorsum, lower lip
p	Lower lip	Tongue tip, tongue dorsum
s	Tongue tip, tongue dorsum	Lower lip
t	Tongue tip	Tongue dorsum, lower lip
v	Lower lip	Tongue tip, tongue dorsum
z	Tongue tip, tongue dorsum	Lower lip

The situation is more complicated for vowels, because the critical gestures for vowel production do not rely on constriction in a single narrow region in the vocal tract but on multiple regions or at least on a wider constriction compared to consonants (Jackson and Singampalli, 2009; Recasens *et al.*, 1997). For example, it is not clear which tongue sensor is most representative, and how much, for the wide palatal region of most vowels. Also it is not straightforward how to choose sensors for a pharyngeal constriction gesture critical for /æ/, /ɑ/, and /ɔ/. Although Jackson and Singampalli (2009) suggested a simple way (i.e., tongue-tip sensor for front, tongue-blade sensor for mid and tongue-dorsum sensor for back vowels) to achieve this, there are still other issues, such as the inter-subject variability in terms of vocal tract shape and articulatory controls or the inconsistency of attached sensor positions across speakers at data collection. For these reasons, we analyze vowels separately from consonants and compare the relative articulatory variability within a relevant articulator for the production of the vowels. For example, the palatal constriction gesture is critical for /i/, hence we consider the vertical movement of the tongue dorsum to be more critical than the horizontal movement for the vowel. Because the pharyngeal constriction gesture is critical for /ɑ/, we consider the horizontal movement of the tongue dorsum to be more critical than the vertical movement.

IV. LANDMARKS-BASED ANALYSIS ON SYLLABLE SEGMENTS

In this section, we investigate how emotion affects articulatory kinematics during syllable production, conditioned on whether the articulators in question are deemed linguistically critical for the initial and final consonants in the consonant-vowel-consonant (CVC) syllables considered. Syllable segments allow us to study emotion variation at phone-target points as well as consonant-vowel transitions for different levels of linguistic criticality of articulators. Our hypothesis is that emotion coloring is more prominently manifested in the kinematics of non-critical articulators than in the kinematics of critical articulators. Because inter-speaker differences in emotion expression are widespread but not fully understood, all experiments in this paper were done separately for each speaker of the database (within-speaker analysis). We begin with analyzing how emotion-related variability of articulatory movements in the syllable varies depending on the linguistic criticality of the articulators.

A. Selection of syllables

We chose CVC syllables with identical place of articulation for the first and second consonants for fair comparisons between constricting and releasing articulatory movements. More specifically, we used the syllables, /p i p/, /n aɪ n/, /f aɪ v/, /p ɑ p/, /t aɪ t/, /n aɪ t/, /p aɪ p/, /d oʊ n/, and /p u p/. Table III shows the number of CVC syllable segments used for analysis. The movements of critical articulators for the consonants in these syllables are captured by the tongue tip or the lower lip sensor. Note that we consider the emotional quality of each monosyllabic word the same as that of the corresponding utterance produced by the actors.

TABLE III. Number of CVC syllable samples used for analysis. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness, and Sad is sadness. CAC1 indicates the critical articulator of the first consonant of its syllable, and CAC2 indicates the critical articulator of the second consonant of its syllable. TT is the tongue tip, L is the lips.

			JN					JR					SB				
CVC	CAC1	CAC2	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad	Neu	Han	Can	Hap	Sad
/p i p/	L	L	4	4	15	7	6	7	3	14	11	8	9	4	9	8	8
/n aɪ n/	TT	TT	20	33	29	19	22	32	35	15	20	34	29	26	19	18	24
/f aɪ v/	L	L	14	22	22	10	14	22	30	8	10	24	20	18	14	10	16
/p ɑ p/	L	L	6	6	10	10	8	10	2	13	11	9	8	7	7	7	8
/t aɪ t/	TT	TT	6	11	7	9	8	10	5	7	10	10	9	8	5	8	8
/n aɪ t/	TT	TT	10	14	22	20	14	16	8	18	22	20	18	12	16	12	16
/p aɪ p/	L	L	6	11	7	9	8	10	5	7	10	10	9	8	5	8	8
/d oʊ n/	TT	TT	2	5	16	9	10	10	12	9	4	14	9	6	9	5	11
/p u p/	L	L	5	7	11	10	7	8	4	9	11	10	9	6	8	6	8

B. Extraction of articulatory parameters

We targeted five linguistically important landmark time points in each CVC: (1) Onset of the release of the first consonant, (2) instant of maximum velocity during the release of the first consonant, (3) instant of maximum opening of the vowel, (4) instant of maximum velocity during the movement toward the second consonant closure, (5) onset of the second consonant closure. The landmark points were selected based on the vertical trajectory of the critical articulator (for onset/coda consonants). Figure 2 illustrates where the five landmark points are located on the lower lip trajectory on the vertical axis during a /p ɑ p/ syllable segment. Landmarks 1 and 5 for each syllable were determined in the vertical movement range, algorithmically chosen to be at the position 3% lower from the highest position value in the CV regions and of the VC regions, respectively. Landmark 3 was determined at the maximal vertical displacement of the jaw from the occlusal plane in the vocalic region. Landmarks 2 and 4 were determined at the points of the maximal absolute first-order derivatives in the CV regions and in the VC regions, respectively. The landmarks serve as basis for the standardization of articulatory parameters of different speakers and emotions.

At these five landmarks, articulatory kinematic parameters (position, speed, and acceleration) of critical and non-critical articulators were measured. More specifically, positions in the horizontal and vertical directions, tangential speed, and tangential acceleration were extracted from EMA sensor trajectories at each landmark. To minimize the effects of sensor tracking error, CVC syllables were discarded if any extracted position parameter was outside the $\pm 3\sigma$ range from the mean of the parameter, where σ and the mean were calculated over all data for each speaker. In total, 20 kinematic parameters (5 landmarks $\times$ 4 measurements) were extracted for each articulator.

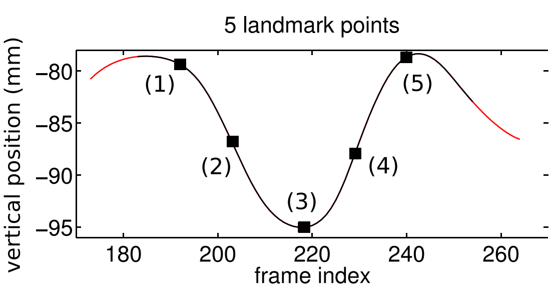

FIG. 2. (Color online) The five landmark points on the trajectory of the lower lip in /p a p/.

C. Statistical analysis of syllable-level kinematics

In this section, we investigate what kinematic aspects of critical and non-critical articulators reveal significant emotional information and how they differ as function of criticality. A Kruskal–Wallis test (Kruskal and Wallis, 1952) was conducted on each articulatory kinematic parameter to reveal which parameters are emotionally significant. The null hypothesis of this test is that the samples (of each parameter) come from the same populations (of categorical emotions), while the alternative hypothesis is that the samples comes from populations, at least two of which differ with respect to location. The Kruskal–Wallis test is a nonparametric method that does not assume normal, or Gaussian, distribution of data points. We observed that sample distributions vary depending on articulatory parameters and that some features do not have normal distribution. For example, the distribution of tongue tip position in the x axis at landmark 1 (onset of release) was close to the normal distribution, while the distribution of tangential speed of the tongue tip at landmark 3 was considerably skewed to zero. Note that tangential speed at largest opening point is supposed to be small in general. The result of a Shapiro–Wilk test on the tangential speed of the tongue tip at landmark 3 of each emotion and each speaker also supported that the population of the feature was not normally distributed ($p < 0.05$).

Figure 3 shows the p value of the Kruskal–Wallis test on each parameter for the five emotion classes. This figure shows, on top of large speaker dependence in emotional variation reflected in individual articulatory parameters, speaker-independent aspects of emotion-dependent articulatory variability in the given datasets of the three speakers. First, this figure shows that all tangential accelerations of the tongue tip are significantly different among five emotions for all speakers [$H(4) < 9.49$, $p < 0.05$]. This result implies that emotion influences the variation of tangential acceleration of the tongue tip throughout the entire syllable regardless of the linguistic criticality of the tongue tip. Also the vertical positionings of the tongue tip at landmarks 2 and 3 are statistically significant [$H(4) < 9.49$, $p < 0.05$], which indicates that the

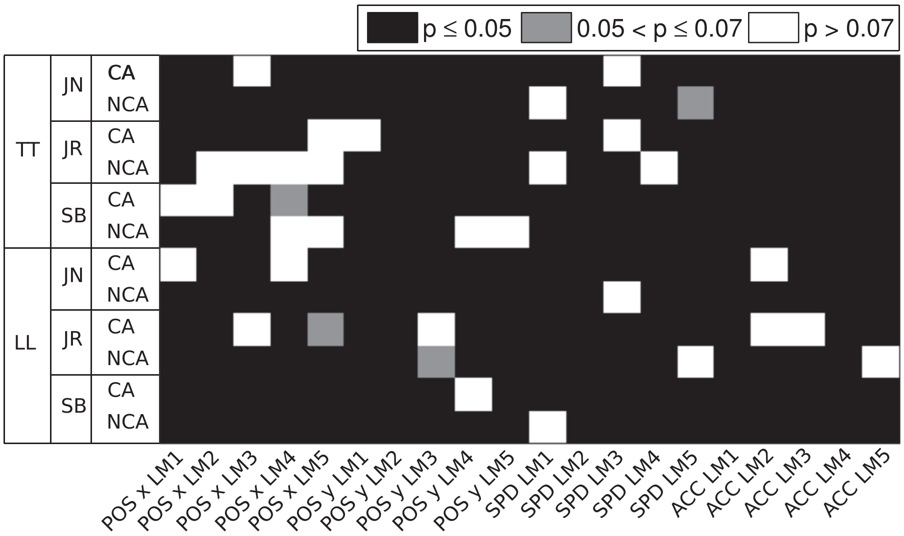

FIG. 3. *P* value of Kruskal–Wallis test on each articulatory parameters, such as horizontal and vertical positions, tangential speed and tangential acceleration at each landmark point. "POS x" is the position in the *x* axis, "POS y" is the position in the *y* axis, "LM" is landmark, "SPD" is tangential speed,"ACC" is tangential acceleration.

vertical positions of the tongue tip during releasing and at the largest opening are affected by emotion expression for all speakers. For lower lip parameters, horizontal position at the landmark 2, vertical position at landmarks 1, 2, and 5, tangential speed at landmarks 2 and 4, and acceleration at landmarks 1 and 4 are statistically significant for all speakers [$H(4) < 9.49$, $p < 0.05$]. The difference of significant parameters of the tongue tip and the lower lip suggests that emotion-dependent variability appearing in articulatory movement margins (for the CVC syllables examined) is articulator-dependent.

Figure 3 also shows that some articulatory parameters of both tongue tip and lower lip are significantly different for the five emotions. Vertical position and tangential speed at the landmark 2 (maximum speed point during constriction release) and tangential acceleration at the landmark 4 (maximum speed point during constriction formation) are statistically significant parameters [$H(4) < 9.49$, $p < 0.05$] for all speakers, indicating that the movements of the two articulators during transition regions between two adjacent linguistic target positions are important sources of emotional information.

It is also observed in Fig. 3 that some kinematic parameters of critical articulators are significant at landmarks 1 and 5 but not significant at landmark 3. For example, for speakers JN and JR, when the tongue tip is critical, tangential speed at the landmark 3 is not statistically significant [$H(4) > 9.49$, $p > 0.05$], while tangential speed at landmarks 1 and 5 is significant [$H(4) < 9.49$, $p < 0.05$]. For a better understanding of this phenomenon, we examined the horizontal and vertical speed at the releasing onset and constriction/closure onset points separately. Figure 4 shows the histograms of vertical velocity for each emotion at releasing onset (landmark 1) in JN's data, as an example. We found that articulatory speed at onsets of release and closure were still significantly affected by emotion. More specifically, on average, higher horizontal and vertical speeds were detected at the two onsets in high arousal emotions, such as hot anger and happiness, than in the other emotions (e.g., sadness), while horizontal speed was lowest in sadness (low arousal emotion). In fact, tangential speed of critical articulators at landmarks 2 and 4 is also statistically significant for both tongue tip and lower lip in all speakers data in Fig. 3. These results indicate that on average, initial and maximum articulatory speed during consonant-to-vowel transition, and maximum and final articulatory speed during vowel-to-consonant transition contain significant emotional information for critical articulators. These trends were not consistently detected in the non-critical articulators.

D. Analysis on the distributions of articulatory positions

We also compared the distributions of the positions of critical and non-critical articulators at landmarks using

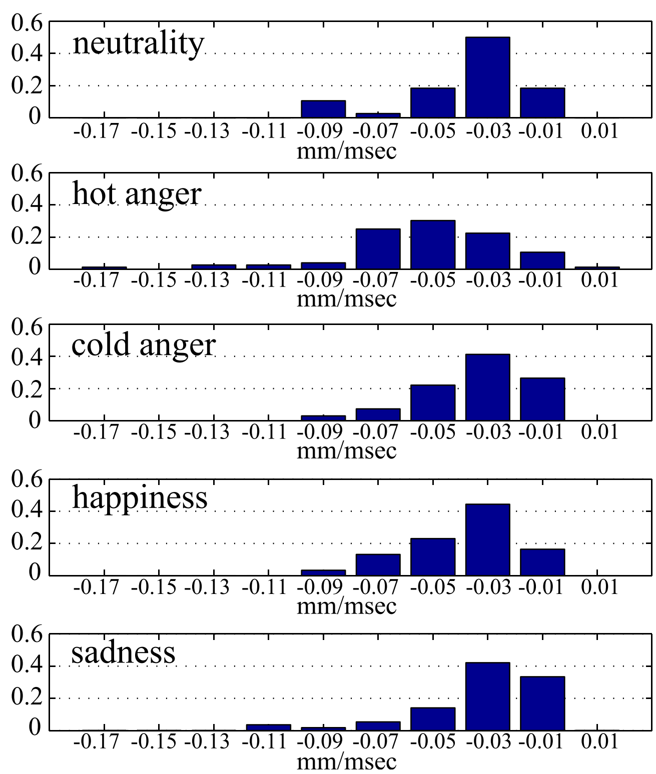

FIG. 4. (Color online) Histograms of the vertical velocity of the tongue tip at releasing onset point (landmark 1) for each emotion in JN's data.

two-sigma (two standard deviations) ellipses. Our goal with this analysis was to understand in what ways the movements of critical and non-critical articulators vary across emotions. Figure 5 illustrates emotion-dependent articulatory variability depending on the linguistic criticality. It is noted that the contrast of criticality of the tongue tip and the lower lip is greatest for consonants (landmarks 1 and 5), and the contrast decreases as being closer to the largest opening (corresponding to landmark 3) for vowels, because the criticality of articulator is categorized for consonants not for vowels. Considering only the case of critical articulators (subfigures in the first and third rows of Fig. 5), we observe that the divergence of non-neutral emotion ellipses from a neutrality ellipse was associated with the arousal dimension of emotions when the articulators were critical (for consonants), i.e., high arousal emotions showed greater divergence

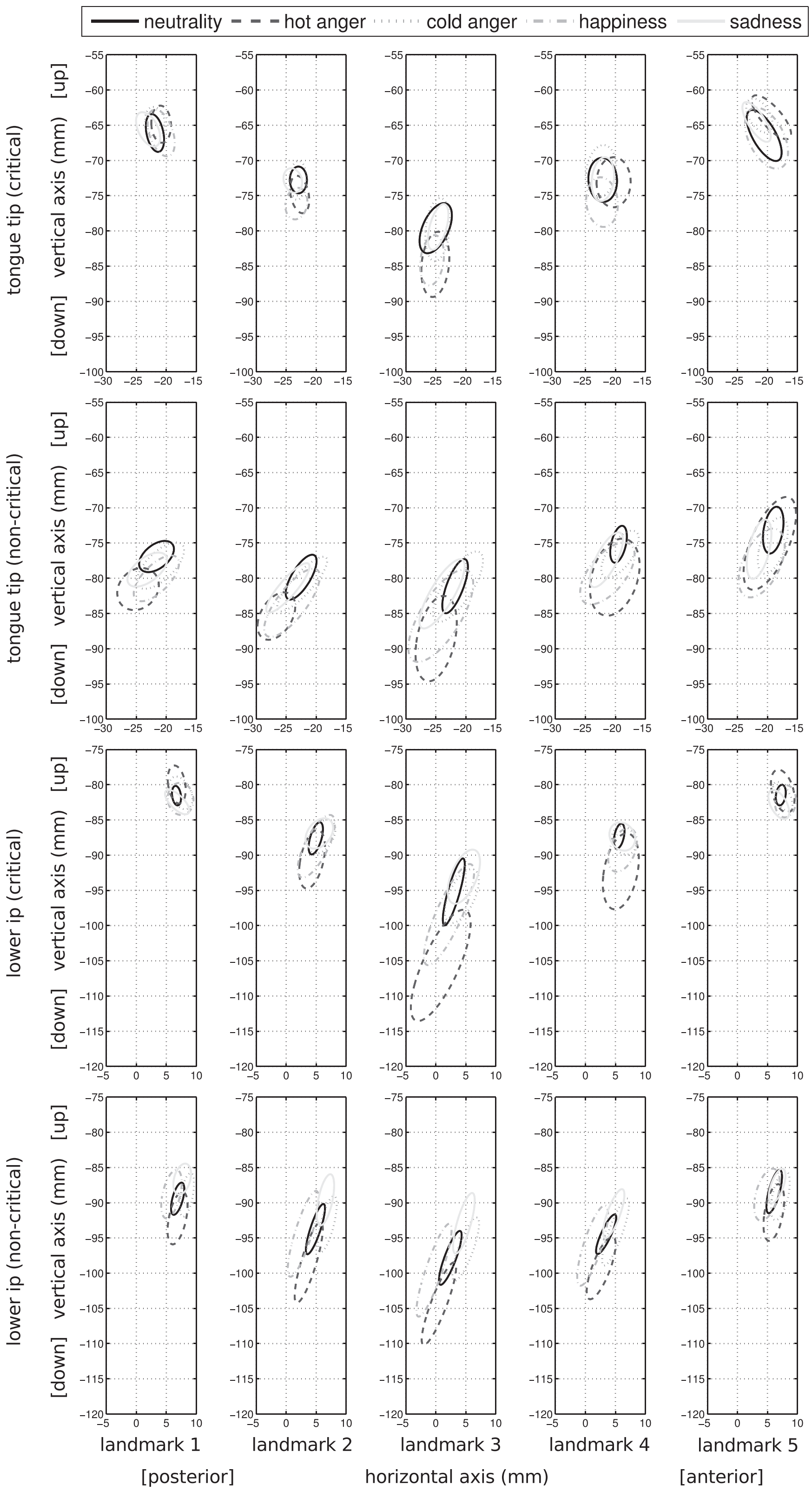

FIG. 5. Example plots (speaker JN) of sample distributions (represented by two-sigma ellipses) of articulatory positions at different landmarks.

from neutrality than low arousal emotions. The dispersion of ellipse centers across emotions tended to be maintained throughout the CVC syllable regions for non-critical articulators. For example, when the lower lip is critical, the center of the hot anger ellipse is located higher than the center of the neutrality ellipse at landmark 1, then located lower at landmarks 2 and 3 (releasing and largest opening points, respectively), and finally located higher again at landmark 5 (closure formation). On the other hand, when the lower lip is non-critical, the relative locations of ellipse centers are consistent for all landmarks. This difference between critical and non-critical articulators is observed for all speakers.

Figure 6 illustrates emotional variations on articulatory trajectories of the tongue tip and the lower lip depending on their criticality. The articulatory segment is for "nine tight night pipes" (see Sec. IV A for their sentence stimulus). To compare articulatory trajectories of different emotions, articulatory trajectories of each instance were aligned to a reference trajectory that is arbitrarily chosen from neutral emotion data. Dynamic time warping (Sakoe and Chiba, 1978) with Euclidean distance between articulatory trajectories was used for the alignment. After alignment, average trajectories for each emotion were obtained by computing the frame-level mean of trajectories after a spline interpolation to 400 frames.

It was observed that when the articulator was critical for consonants, the articulations of hot anger and happiness often showed larger openings for vowels and exaggerated constrictions for consonants, while the articulations of sadness often showed smaller openings for vowels and smaller constrictions for consonants. When the articulator was non-critical, relative articulatory positioning among different emotions tended to be less sensitive to closure and opening gestures of the co-occurring critical articulators. This contrast of critical and non-critical cases was also observed in the data of all speakers except the non-critical cases of JR's data. However, the two-sigma ellipses plot (drawn with more data than the four words) of JR's data supports that the relative articulatory positioning of different emotions for non-critical cases varies less in JR's data. This may be due to the speaker-specific characteristic of JR that is the significantly correlated vertical movements of the tongue tip and the lower lip as shown in Fig. 7. This figure shows correlation coefficients of tongue tip and lower lip movements in the vertical direction that were computed from the average vertical trajectories of the two articulators for each emotion in "nine tight night pipes." Figure 7 indicates that the vertical movements of the two articulators are most correlated for all emotions in JR's data. Highly correlated movements of the two articulators of JR imply that the lower lip and the tongue tip movements are being coupled during the production of the monosyllabic words. This may suggest that the dependency between articulators is not a static parameter that is associated with only anatomical structure and coordination for linguistic encoding, but a dynamic parameter related to other factors, e.g., para-linguistic factors.

Figure 5 also shows the difference of articulatory variability during closure-to-releasing and approaching-to-closure motions. If the tongue tip is critical for consonants, the articulatory position shows greater variation at the landmark 1 than at the landmark 5 in terms of ellipse sizes and the dispersion of ellipse centers in the horizontal and vertical directions. In fact, such variation of articulatory position is greater at the landmark 2 than the landmark 4. This is counterintuitive, because constriction formation of critical articulators for consonants is more actively and carefully controlled than releasing, hence approaching motions are likely to show less variability. The effect of co-articulation may be one possible reason for this phenomenon. The constriction gesture of the tongue tip for the second consonants in /n aɪ t/ and /n aɪ n/ may have become loosened by

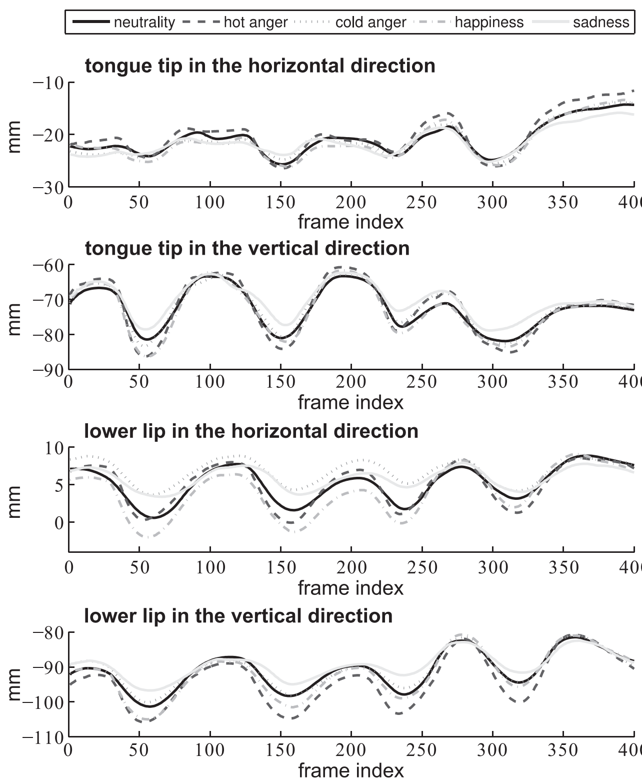

FIG. 6. Averaged articulatory trajectories of each emotion for "nine tight night pipes" in the sentence 4 in JN's data. The two plots from the top show the averaged trajectories of the tongue tip; the other two plots show those of the lower lip.

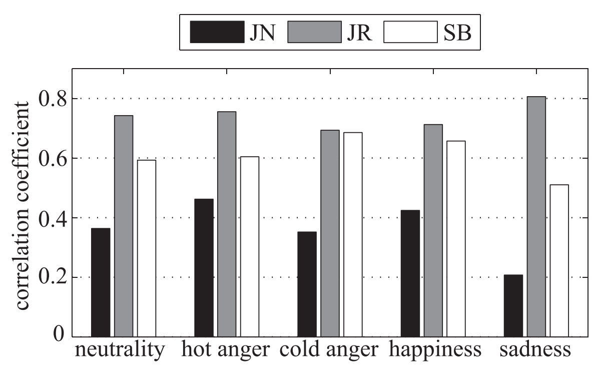

FIG. 7. Correlation coefficients of two averaged trajectories of the tongue tip and the lower lip in the vertical direction.

overlapping closure gestures of the lower lip for the following consonants, bilabial /p/ and labio-dental /f/, respectively (see the Appendix for full sentences).

In summary, we found that emotion affects the kinematics of both non-critical and critical articulators in the CVC syllable segments considered. The emotionally significant kinematic parameters vary depending on speakers and articulators. In addition, when articulators are critical for consonants in the CVC syllables, greater openings for vowels and stronger constrictions for consonants for high arousal emotions are observed. When articulators are non-critical, relative articulatory positioning for different emotions tend to be less sensitive to closure and opening gestures of the co-occurring critical articulator. These results suggest that emotion affects the articulatory positioning for non-critical articulators and the movement range of critical articulators, controlled for achieving (short-term) linguistic goals.

V. ARTICULATORY VARIABILITY ANALYSIS AT THE PHONE LEVEL

Section IV studied emotional variation in articulatory kinematics depending on the linguistic criticality of articulators with a limited dataset of CVC syllables. This section studies emotional variation of articulatory behaviors at phone target positions using the entire EMA database (excluding fast style speech as noted earlier).

A. Experimental Setup

The comparison of critical and non-critical articulators in terms of the articulatory variability of target phone position for different emotions requires determining when articulators reach the target position (steady-state point) for each phone. Manual selection of steady-state points in the whole EMA database is time-consuming. Acoustic boundaries determined by an automatic phonetic alignment do not inform articulatory target points directly. We determined the best steady-state point of each phone as follows. First, we determined phonetic boundaries using a hidden Markov model based automatic phonetic alignment toolkit, the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008), followed by manual correction of misalignment outputs. For each phone, we searched for the best steady-state point among multiple candidates. In the case of vowels, candidate points comprised the articulatory positions on the x and y axes of each articulator at three points: The middle frame, 20 ms before the middle frame, and 20 ms after the middle frame with respect to the manually corrected phone boundary. In the case of consonants, candidate point comprised the articulatory positions on the x and y axes of each articulator at the same three points, plus four additional points: (a) The highest point on the y axis in a large marginal region (20 ms before and after phone boundary); (b) the highest point on the y axis in a small marginal region (10 ms before and after phone boundary); (c) minimum tangential speed points in the large marginal region; (d) minimum tangential point in the small marginal region. The reason for using the marginal regions instead of just phone boundaries is that acoustic phone boundaries do not always include the steady-state points of articulatory movements. Next, for each phone, we calculated the mean of Euclidean distances from the median sample point to each articulatory position sample of each of the three frames for vowels or each of the seven frames for consonants. The frame of the least mean value was selected as the steady-state point for the phone.

Our foremost interest is the behavior at the phone level of (linguistically) critical and non-critical articulators in emotional speech production. However, from a statistical experimental design perspective, the phone identity factor is nested in the criticality factor, which means that any analysis interested in the articulatory variability for different degrees of linguistic criticality requires a normalized representation of the variability across phones. To address this problem, we tried two different methods for parameterizing articulatory variability fairly applicable across phones, using (1) the centroid distances between emotion cluster pairs and (2) the mean deviation within emotion for every phone. The details of the two methods are explained in the sections that follow.

B. Analysis of inter-emotion variability of articulatory kinematics

In this section, we investigate how the degree of linguistic criticality of articulators is associated with inter-emotion variability. In particular, we examine which articulatory type (critical or non-critical) displays more inter-emotion variability in the articulatory phonetic target position. The inter-emotion variability in articulatory target positioning is quantified by the average of the centroid distances between emotion cluster pairs, where the centroid is the mean position of all samples of each emotion cluster. This parameter measures the averaged distance between the centers of different emotion clusters for each phone. Let x_i^k denote the arithmetic mean of all samples of the vertical or horizontal coordinate of a given EMA sensor's position in emotion i and phone k. Suppose the number of emotion clusters is N. Then the average of the centroid distances between emotion cluster pairs of phone k, denoted by D_k, is calculated as

$$D_k = \frac{1}{C(N,2)} \sum_{i,j} d\left(x_i^k, x_j^k\right), \quad i \neq j, \tag{1}$$

where $C(N, 2)$ is the number of two-combinations from N elements; $d(a, b)$ is the Euclidean distance between a and b.

Figure 8 shows the box plots of D_k of critical and non-critical cases for each articulator. First, this figure shows that on average, the D_k of the horizontal lower lip positions at phone targets is greater for non-critical articulators than for critical articulators. We also conducted a one-tailed t-test with the hypothesis that the D_k of the horizontal lower lip positions at phone targets is greater for non-critical cases than for critical cases. Results indicated that the difference between critical and non-critical cases was significant for JR ($t = 3.31$, $p = 0.00$) and SB ($t = 2.68$, $p = 0.01$), but it was not significant for JN ($t = 1.43$, $p = 0.09$) at the 0.05 level. For the vertical position of the lower lip, the difference between critical and non-critical cases in terms of D_k was speaker-dependent: The average of D_k was greater for

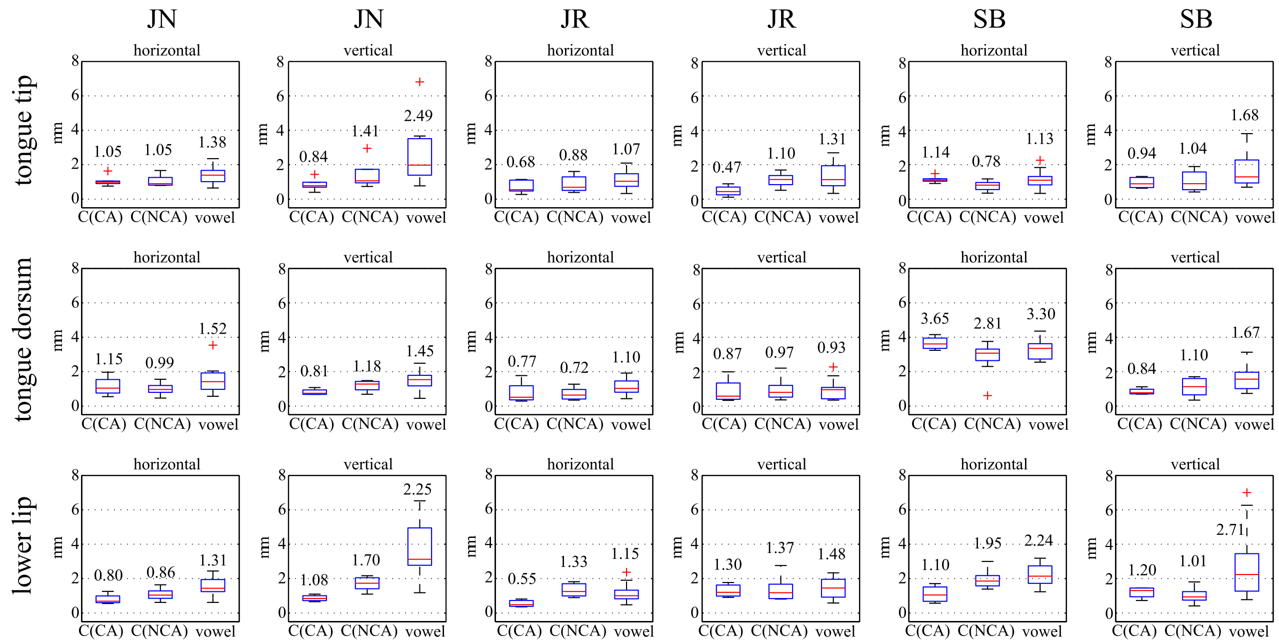

FIG. 8. (Color online) Box plots of the average of centroid distance among emotion cluster pairs. CA is critical articulator case, NCA is non-critical articulator case. C() denotes consonants. Vowels are analyzed separately from consonants because of their different nature for determining critical or non-critical articulator in this study. The value above each box plot is the mean of each case.

non-critical cases than for critical cases in the data of JN and JR, while it was the other way round in SB's data. These results suggest that on average, the mean positions of the lower lip for different emotions are more consistently dispersed for non-critical cases compared critical cases, only in the horizontal direction.

We investigated the variation of the horizontal position of the lower lip more specifically for each emotion and each phone. Figure 9 shows the mean horizontal position of the lower lip of each emotion after aligning the mean of neutrality to be 0 for each phone in SB data as an example. We found that the lower lip showed a posterior position for most phones when speakers expressed happy emotions than the other emotions. These trends were observed in the data of all speakers. The retraction of the lower lip for happiness was statistically significant in the result of one-tailed t-test ($t = 7.38$, $p = 0.00$). We also found that on average, the retraction of the lower lip for happiness occurred more significantly for non-critical cases than for critical-cases. One possible reason is that when speakers express happy emotion, their lip might have been stretched to the sides often, pulling the lower lip backward (smile-like gesture).

For each of the tongue sensors, the average D_k of the vertical position was larger for non-critical cases than for critical cases for all speakers. This indicates that on average, the vertical position of the tongue tip and the tongue dorsum at phone targets is more dispersed for non-critical cases than for critical cases. The difference between non-critical and critical cases was statistically significant at the 0.05 level for the tongue dorsum data of JN ($t = 2.33$, $p = 0.02$) and the tongue tip data of JR ($t = 0.96$, $p = 0.01$) by one-tailed t-test, while it was not significant for the other cases. On the other hand D_k of horizontal position was (even slightly) greater for critical cases than for non-critical cases for all speakers' data. The difference, however, was not significant for any of them at the 0.05 level of the one-tailed t-test. These results indicate that on average, the mean positions of the lower lip for different emotions are more dispersed for non-critical

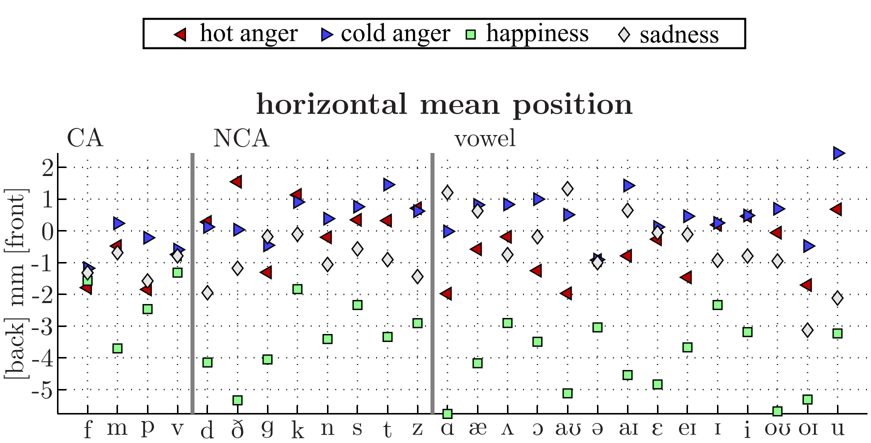

FIG. 9. (Color online) Relative mean (centroid) of the horizontal position of each emotion to the neutrality (which is aligned to 0 on the y axis) in SB's lower lip data.

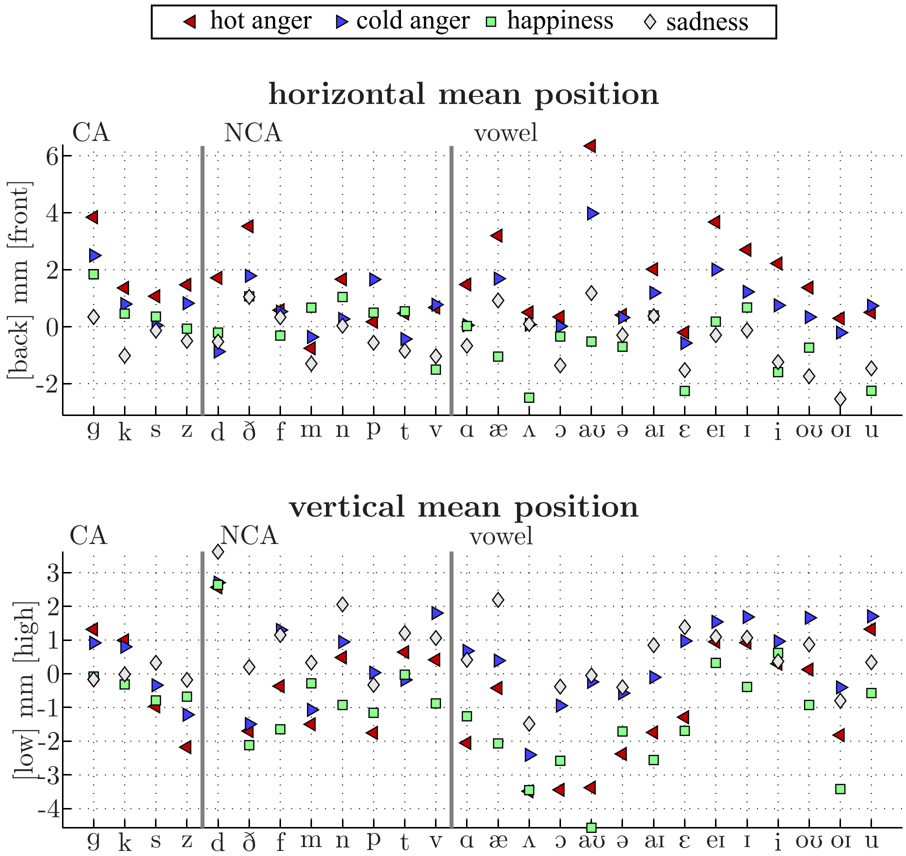

FIG. 10. (Color online) Relative mean (centroid) of the horizontal (top subplot) and vertical (bottom subplot) positions of each emotion cluster for each phone to the neutrality (which is aligned to 0 on the y axis) in JN's tongue dorsum data.

cases than critical cases, consistently only in the vertical direction.

From our investigation on the tongue data for each emotion and each phone, it was observed that on average, the tongue dorsum as a critical articulator showed more upward constriction for velar stops, such as /g/ and /k/, when the emotional state of speakers were hot anger (and cold anger for JN and SB) as opposed to other emotions. More forward positioning of the tongue dorsum was also observed for hot anger, except for /k/ in SB's data. Figure 10 shows the horizontal and vertical positions of the tongue dorsum for each emotion and each phone in JN's data as an example. These results suggest that tongue dorsum closure gesture is stronger for hot anger than for the other emotions.

Finally we also examined whether the tongue tip and the tongue dorsum showed different emotional variance depending on the linguistic criticality for vowels. According to the gestural description in the task dynamics application model (Nam *et al.*, 2004), palatal constriction gesture is critical for /i/, /ɪ/, and /ɛ/, while pharyngeal constriction gesture is critical for /æ/, /ɑ/, and /ɔ/. It is reasonably assumed that the linguistic criticality of the vertical tongue dorsum position is higher for the palatal vowels than for the pharyngeal vowels. Also it is assumed that the criticality of the horizontal tongue dorsum position is higher for the pharyngeal vowels than the palatal vowels. Figure 11 shows the average of D_k of the tongue dorsum position in the horizontal and vertical directions for each of palatal and pharyngeal vowels. We found that on average, D_k of the tongue dorsum horizontal position was greater for the palatal vowels than for the pharyngeal vowels in all speakers' data, while D_k of the tongue dorsum vertical position was greater for the pharyngeal vowels than for the palatal vowels. This result implies that for vowels, the tongue dorsum position in less constrained direction displays more inter-emotion variation than in more constrained direction.

C. Analysis of within-emotion variability of articulatory kinematics

In Sec. V B, we investigated inter-emotion variability of the mean position of articulators as a function of linguistic

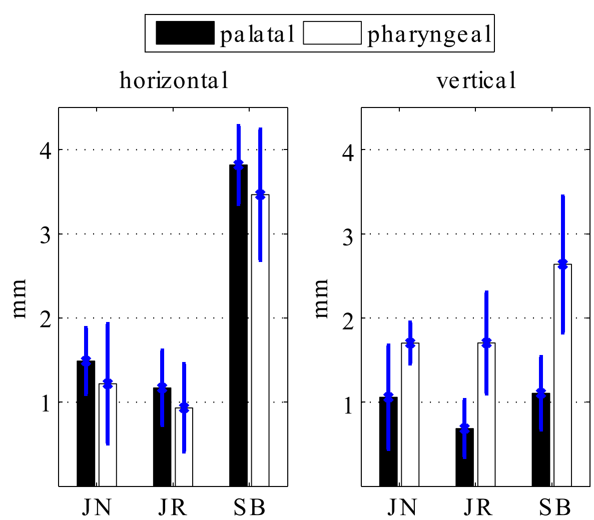

FIG. 11. (Color online) The average of centroid distances of emotion cluster pairs for the horizontal (left plot) or vertical (right plot) tongue dorsum positions, contrasted based on critical constriction gestures, such palatal constriction and pharyngeal constriction of the tongue dorsum, for vowels.

TABLE IV. The results of one-tailed t-test with mean deviation measure on the hypothesis that non-critical articulator has greater range of articulatory target position for each phone within emotion than critical articulator. "x axis," "y axis" indicate the results of test on mean deviation value of the horizontal articulatory position or the vertical articulatory position, respectively. Neu is neutrality, Han is hot anger, Can is cold anger, Hap is happiness, and Sad is sadness. Only consonants are included in this analysis. Numbers in bold are statistically significant ($p < 0.05$). t-statistic is out of parenthesis in each cell, and p value is in parenthesis.

		Tongue tip		Tongue dorsum		Lower lip	
Speaker	Emo	x axis	y axis	x axis	y axis	x axis	y axis
JN	Neu	0.46 (0.33)	0.30 (0.38)	0.42 (0.34)	1.49 (0.08)	1.32 (0.11)	**3.88 (0.00)**
	Han	1.13 (0.14)	1.84 (0.05)	0.35 (0.37)	**2.12 (0.03)**	**3.08 (0.01)**	**2.90 (0.01)**
	Can	−1.14 (0.86)	1.66 (0.06)	1.14 (0.14)	**2.35 (0.02)**	1.84 (0.05)	**4.85 (0.00)**
	Hap	−0.48 (0.68)	0.62 (0.27)	**2.39 (0.02)**	1.45 (0.09)	**2.48 (0.02)**	**2.68 (0.01)**
	Sad	−0.44 (0.66)	0.47 (0.32)	1.64 (0.07)	**2.49 (0.02)**	1.75 (0.06)	**2.89 (0.01)**
JR	Neu	1.34 (0.11)	**2.44 (0.02)**	**2.41 (0.02)**	**3.48 (0.00)**	0.52 (0.31)	1.12 (0.15)
	Han	1.87 (0.05)	**2.13 (0.03)**	**2.11 (0.03)**	1.87 (0.05)	**2.13 (0.03)**	1.77 (0.05)
	Can	1.78 (0.05)	**2.44 (0.02)**	**2.78 (0.01)**	**2.04 (0.04)**	**2.24(0.03)**	**2.79 (0.01)**
	Hap	0.44 (0.34)	1.38 (0.10)	**3.34 (0.00)**	**2.31 (0.02)**	1.48 (0.09)	−0.14 (0.55)
	Sad	**2.08 (0.03)**	**2.55 (0.02)**	1.58 (0.07)	**2.65 (0.01)**	**2.01 (0.04)**	1.42 (0.09)
SB	Neu	−0.03 (0.51)	1.86 (0.05)	−1.29 (0.89)	1.16 (0.14)	**2.61 (0.01)**	**2.05 (0.03)**
	Han	−0.01 (0.50)	**1.92 (0.04)**	0.92 (0.19)	1.71 (0.06)	**2.11 (0.03)**	**2.72 (0.01)**
	Can	0.07 (0.47)	1.85 (0.05)	−0.03 (0.51)	**2.48 (0.02)**	**2.31 (0.02)**	**2.59 (0.01)**
	Hap	0.19 (0.43)	1.82 (0.05)	**2.09 (0.03)**	1.44 (0.09)	**2.27 (0.02)**	1.57 (0.07)
	Sad	0.03 (0.49)	**2.04 (0.03)**	0.73 (0.24)	1.27 (0.12)	**1.99 (0.04)**	1.32 (0.11)

constraints. In this section, we investigate within-emotion variability in terms of the range of articulatory positions for critical and non-critical articulators. That is, we study in what way criticality associates with the range of articulatory positions for phone targets in each emotion. We also test our hypothesis that non-critical articulators have more emotional variation, particularly in terms of the within-emotion dispersion at articulatory target points, than critical articulators. We quantified the articulatory range variability of each emotion by the mean deviation of articulatory position samples in the horizontal or vertical direction.

First, we conducted one-tailed t-tests on the hypothesis that non-critical articulators had larger range of articulatory target position than critical articulators for each emotion. Table IV shows the results. We found that on average, non-critical articulators show greater mean deviation than critical articulators in most cases (as indicated by positive t-statistic value in Table IV), which supports the hypothesis, overall. The statistical significance in Table IV is largely speaker-dependent. For example, the difference between critical and non-critical in the vertical mean deviation of the tongue tip data of JN is not significant for any case, while the difference in the vertical mean deviation of the tongue tip data of JR is significant for four emotions (neutrality, hot anger, cold anger, and sadness).

We also examined emotional variation of the positions of articulators depending on linguistic criticality more specifically for each phone. Figure 12 shows the scatter plot of the mean deviation of articulatory target positions in the vertical direction for each phone in SB's data as an example. We observed that overall, the mean deviations of the tongue tip, the tongue dorsum, and the lower lip at phone targets were larger when these articulators were non-critical than when they were critical. However, it is not always true for some cases. For vertical tongue tip position, the mean deviations of velar stops (/g/ and /k/) tended to be lower than those of the other non-critical cases (e.g., /f/, /m/, /p) and even similar to those of critical cases (e.g., /d/, /ð/, /n/). We note that two tongue sensors were placed about 1.5 – 2 cm closer to each other than the anatomical tongue tip and velar closure point on the tongue surface (tongue dorsum), which may have increased the dependency between the tongue tip and tongue dorsum motions. Hence we speculate that tongue tip position was associated with tongue dorsum position more than the lower lip position, resulting in more limited tongue tip position variation for the velar stops than for the labial consonants. We also observe that in the critical cases of the tongue tip, the maximal mean deviations in both horizontal and vertical directions for alveolar fricatives (/s/ and /z/) were always smaller than those for alveolar stops (/d/, /n/, /t/) in all speakers' data, presumably because alveolar fricatives require more careful maneuver than alveolar stops (Subtelny and Oya, 1972).

For the tongue dorsum, closure gesture for velar stops (/g/ and /k/) showed less mean deviation in the vertical direction compared to alveolar fricatives (/s/ and /z/) across all emotions. Although wide opening gesture of the tongue dorsum (as well as constriction gesture of the tongue tip) is essential for a production of the alveolar fricatives (Nam *et al.*, 2004), the result suggests that this wide opening gesture does not require a strict control of the constriction degree. Speaker-independent pattern contrasting the alveolar stops and the alveolar fricatives was not observed in the results of the tongue dorsum horizontal position.

VI. SIMULATION EXPERIMENT

In this section, we examine whether the large variability of non-critical articulators is the mechanical outcome of the controls of critical articulators. We first synthesize non-critical articulatory trajectories on the basis of the physiological

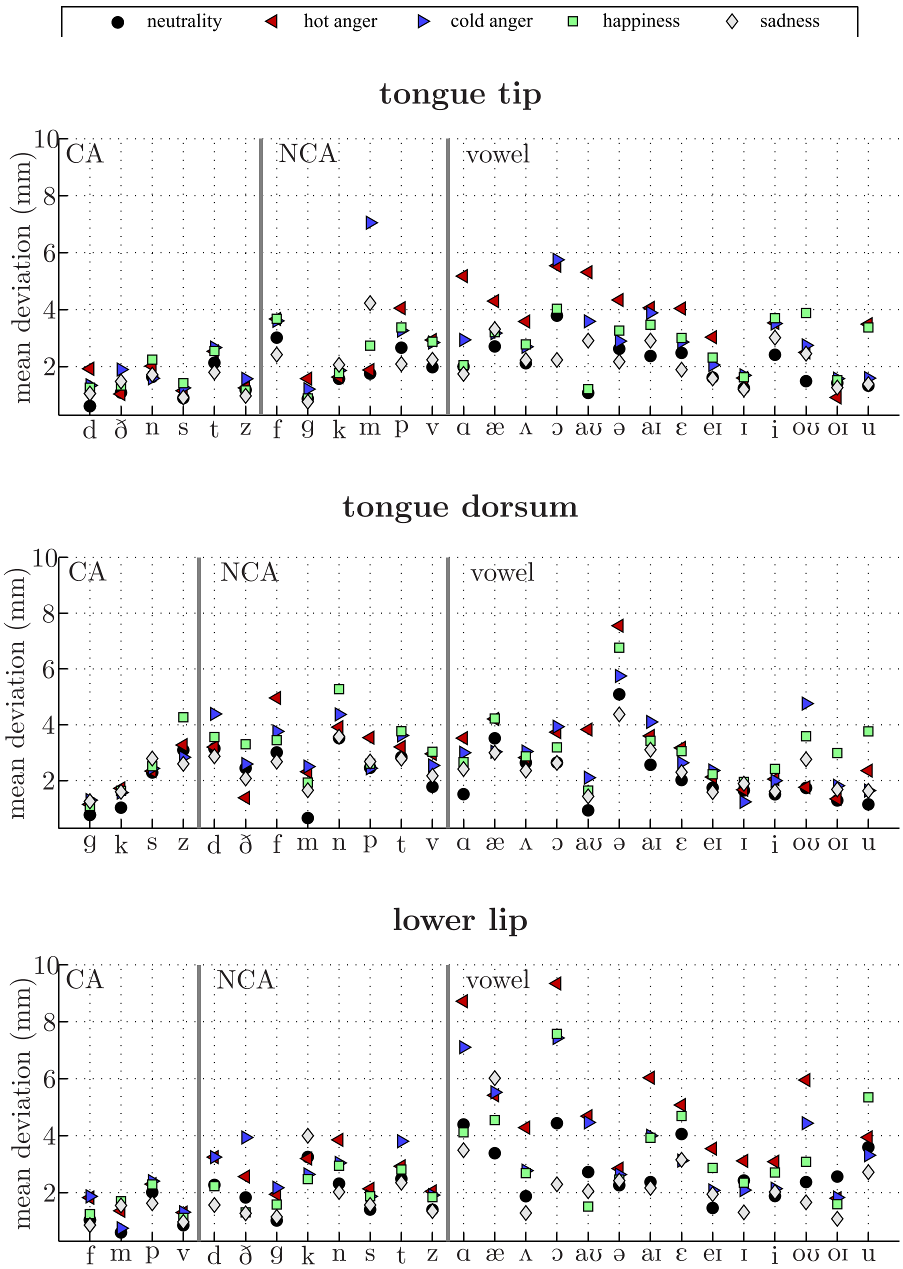

FIG. 12. (Color online) Scatter plots of the mean deviation of articulatory positions of each emotion of SB. Divided by two gray solid lines, the left-most block is critical articulator (noted as CA), middle block is consonant non-critical articulator (NCA), the right-most block is vowel non-critical articulator.

constraints that govern the spatiotemporal relationships among all articulators and the time points when given articulators are linguistically critical. The emotional variations in the synthesized trajectories are compared to the emotional variation in the true data. If the two trajectories of non-critical articulators are similar in terms of emotional variation, it can be inferred that the emotion-dependent variability of non-critical articulators is a secondary effect of the control of critical articulators.

A. Model description

This section describes an articulatory model (Kim *et al.*, 2014b) that was used for the aforementioned simulation experiment. This model estimates trajectories of non-critical articulators based on only the following two factors: (a) The contextual constraints of the preceding or following time points when said articulators are critical, (b) physiological constraints on said non-critical articulators from articulators that are critical at the time point in question. This estimation problem is formulated as the following: Let $f_i(t)$ denotes the position of ith articulator at time t. $f_i(t_c)$ is the position of the i-articulator at the nearest critical time point t_c from the current time t for the ith articulator. Hence this term represents the influence of the contextual constraints from the nearest critical point. $\hat{f}_i^p(t)$ represents the influence of the physiological constraints on the i-articulator. The estimated position of the ith articulator, $\hat{f}_i(t)$ is modeled by convex combination of $f_i(t_c)$ and $\hat{f}_i^p(t)$, using a weighting function $K_i(t) \in [0, 1]$ as follows:

$$\hat{f}_i(t) = f_i(t_c)K_i(t) + \hat{f}_i^p(t)(1 - K_i(t)). \quad (2)$$

The weighting on the contextual factor should be negatively correlated to $|t - t_c|$, but the nature (linear or non-linear) of this function is unknown. Hence $K_i(t)$ is modeled by the non-linear function as follows:

TABLE V. The number of utterances selected for simulation experiment.

Speaker	Neutrality	Hot anger	Cold anger	Happiness	Sadness
JN	30	43	77	53	50
JR	62	44	55	48	67
SB	48	44	46	43	48

$$K_i(t) = \frac{1}{1 + \exp(-\eta(\lambda_i(t) - \xi))}. \quad (3)$$

This sigmoid function can also be close to linear depending on the hyper-parameters η and ξ which are tuned on the development set. $\lambda_i(t) \in [0,1]$ denotes a monotonically increasing function of $|t - t_c|$, thus $K_i(t)$ is monotonically decreasing.

$\hat{f}_i^p(t)$ is a function of the positions of *only* corresponding critical articulators at t as follows:

$$\hat{f}_i^p(t) = \sum_{\substack{l=1 \\ l \neq i}}^{N_C(t)} (\alpha_{i,l} f_l(t)) + \beta_i, \quad (4)$$

where $N_C(t)$ is the number of the corresponding critical articulators at t; $\alpha_{i,l}$ and β_i are the coefficients of the model. It is reasonably assumed that the effect of physiological constraints among articulators can be represented by an affine map. For example, the physiological influence from the position of the jaw to the position of the lower lip is computed by rotation, scaling, and translation, those are affine transformation. Note that the critical articulators' data used for representing $f_i^p(t)$ do not include the data of the ith articulator itself.

Finally, the optimal $\hat{f}_i(t)$ is found by minimizing $\mathcal{J}$,

$$\mathcal{J} = \sum_{t=1}^{M} |f_i(t) - \hat{f}_i(t)|^2, \quad (5)$$

where M is the number of articulatory frames used for tuning the parameters of $K_i(t)$.

B. Synthesis of non-critical trajectories

To minimize the effect of erroneous articulatory data, we excluded utterances in which any of the articulatory data is out of empirically selected upper and lower boundaries. For each dimension of each speaker's data, the upper boundary is 0.95 quantile $+ 2 \times$ standard deviation, while the lower boundary is 0.05 quantile $- 2 \times$ standard deviation. Then each dimension of articulatory data of each speaker is scaled to the range of [0,1] for fair evaluation. Table V shows the number of utterances selected for the simulation experiment.

The critical time point for each critical articulator for consonants is selected at the maximum constriction point of the articulator. We followed the distinction of critical and non-critical articulators in Table II for consonants. For vowels, the critical time point is decided based at the maximum opening point of the jaw and the tongue dorsum (in the vertical direction). The upper lip data are excluded in this experiment because the upper lip is not anatomically constrained to any of the other articulators monitored in this dataset. Also it was reported that the upper lip data did not improve the estimation performance (Kim *et al.*, 2014b).

Our model is trained in leave-one-utterance-out setup for each emotion and each combination set of critical articulators, except the estimating articulator, because $\alpha_{i,l}$ and β_i of f_j^p in Eq. (4) depends on the combination. After the train and development sets are equally divided, $\hat{f}_i^p(t)$ is trained on the train set. The parameters of $K_i(t)$ are tuned on the development set. The performance of our final model $\hat{f}_i(t)$, $\forall i$ is evaluated in terms of the mean of the root-mean-squared-error (RMSE), denoted by E_{RMSE}, and the mean of the correlation coefficient, denoted by E_{CORR}, between the true trajectory and the estimated trajectory of all utterances.

We first demonstrate that our model can estimate articulatory trajectories well for all emotions. Table VI shows the evaluation results of the estimated articulatory trajectories in terms of RMSE and correlation coefficient. Our model shows satisfactory estimation performance (maximum $E_{RMSE} = 0.079$, minimum $E_{CORR} = 0.794$) for data of all speakers and all emotions. This estimation performance is similar to one reported in our previous study (Kim *et al.*, 2014b), which was performed with different sentences and emotions (neutrality, anger, happiness, sadness, and fear). The result in Table VI suggests that the trajectories of non-critical articulators are estimated reasonably well for the five target emotions in the dataset. It also suggests that the positions of non-critical articulators are considerably dependent on the positions of the corresponding critical articulators and the closest critical moment of the (non-critical) articulators. Figure 13 illustrates true and estimated vertical trajectories of the tongue tip, the tongue dorsum, and the lower lip for the sentence "I saw nine tight night pipes in the sky last night," showing high similarity between the trajectories.

TABLE VI. The results of evaluation of the estimated articulatory trajectories. The mean of RMSE or correlation coefficient is shown without parenthesis. The standard derivation is shown in parenthesis.

Speaker		Neutrality	Hot anger	Cold anger	Happiness	Sadness
JN	E_{RMSE}	0.079 (0.030)	0.074 (0.018)	0.074 (0.021)	0.072 (0.017)	0.071 (0.022)
	E_{CORR}	0.848 (0.116)	0.847 (0.096)	0.845 (0.122)	0.855 (0.102)	0.846 (0.130)
JR	E_{RMSE}	0.071 (0.023)	0.070 (0.022)	0.072 (0.024)	0.070 (0.020)	0.069 (0.023)
	E_{CORR}	0.869 (0.121)	0.864 (0.109)	0.856 (0.125)	0.865 (0.112)	0.862 (0.117)
SB	E_{RMSE}	0.079 (0.035)	0.071 (0.026)	0.077 (0.023)	0.071 (0.026)	0.075 (0.021)
	E_{CORR}	0.847 (0.135)	0.794 (0.141)	0.814 (0.116)	0.845 (0.119)	0.801 (0.162)

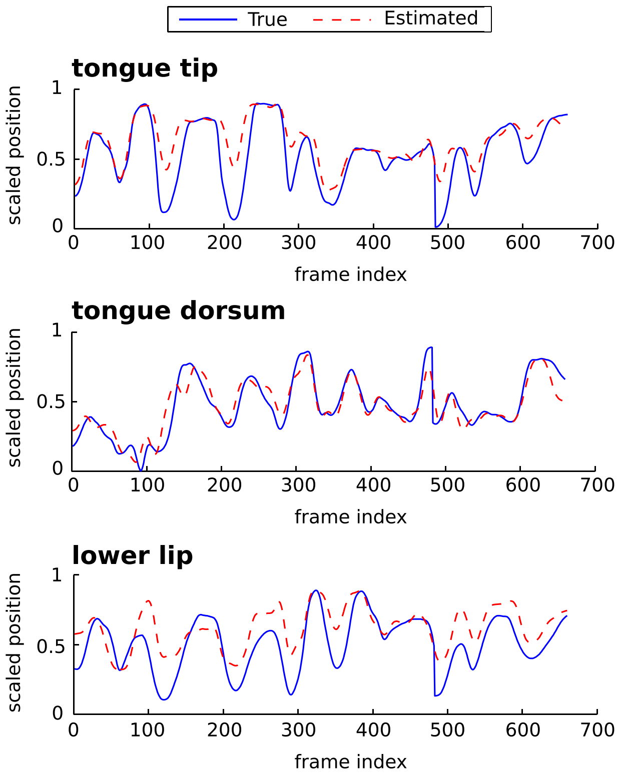

FIG. 13. (Color online) Example plots of the true and estimated trajectories of the tongue tip, the tongue dorsum, and the lower lip in the vertical direction. An utterance of neutral emotion in JN's data is used.

C. Analysis

In this section, we discuss the emotion-dependent variability of the non-critical articulators by comparing the true and simulated articulatory data. The emotion-dependent variability between the true articulatory data and the estimated data from the aforementioned model was compared by means of discriminant analysis and a statistical test. For the discriminant analysis experiment, emotion model was trained on the true data and tested in the estimated data. The classification accuracy on the estimated data was compared to the classification accuracy on the true data. Similar accuracy between the two results is an evidence for the high similarity between the true and estimation data in terms of emotion-dependent articulatory variability. For discriminant function, we used 2D normal density model, one mode for each emotion, in the Mahalanobis distance space.

The test statistic of the pair-sample t-test was used for the similarity metric of the two distributions: One for the true data and the other for the estimated data. The analysis was performed for each phone, each emotion and each articulator. Each of true and estimated data was subtracted to the centroid of corresponding neutral data so that the distribution of each emotion represents the deviation from the neutral emotion.

D. Results

Figure 14 shows the emotion classification results. In most cases, the classification accuracy of estimated data is similar to the classification accuracy of true data. This indicates high similarity between true and estimated data in terms of the emotion-dependent variation of articulatory position distribution and suggests that the large variability of non-critical articulators depending on emotion is significantly dependent on the controls of critical articulators.

We also investigated the similarity between true and estimated data for *individual* emotions using the pair-sample t-test. The p value less than 0.05 indicates that the means of the two distributions (true and estimated data) are statistically significantly different at the level of $\alpha = 0.05$, suggesting that overall, the emotional variations of the true and simulated data are significantly different in terms of their means. Although high p value cannot be directly interpreted as a *statistical* evidence for the validity of null hypothesis, it can be used as a similarity metric of true and estimated data in terms of their mean. Figure 15 shows the p value of pair-sample t-test for each speaker, each phone, each dimension, and each articulator. Results of neutrality are omitted because the means of their distributions are always 0. Note that the distribution of each emotion was normalized by subtracting to the centroid of neutrality. In many cases (Not marked by an asterisk in Fig. 15), the p value is greater than 0.05, so the null hypothesis that the mean of true and estimated data are significantly different cannot be rejected at the level of $\alpha = 0.05$ for these cases. In fact, the p value is often considerably high. This is a supporting evidence for high similarity of the mean of true and estimated data in the cases. The result of high similarity suggests that the postural variation of non-critical articulators is often significantly dependent on the controls of critical articulators. The cases of high similarity are not consistent across speakers, suggesting large speaker variability on the dependency of non-critical articulators to critical articulators in emotional speech.

VII. DISCUSSION AND CONCLUSIONS

This study provides evidence that the emotional variation pattern of articulatory positions during CVC syllables depends on the degree of linguistic criticality of articulators for the first and the final consonants. When articulators are critical for the consonants in the CVC syllables, high arousal emotions show more peripheral articulatory movements with large movement range, especially in the vertical direction, while it was the other way around for low arousal emotions. The dispersion pattern of critical cases is in line with the experimental results of previous studies, i.e., large movement range and large opening for anger (Lee *et al.*, 2005; Lee *et al.*, 2006; Lee *et al.*, 2008; Erickson *et al.*, 2000). Relative articulatory positioning of the five emotions for non-critical cases is not as sensitive to the manner of articulation for each phone as those for critical cases. One possible implication of these articulatory variations is the modulation of the vocal tract variables in the task dynamics as a result of emotion coloring. For example, when the tongue tip is critical for the initial or final consonants in a CVC syllable, tongue tip constriction degree in vowel regions can be higher (larger opening) for high arousal emotions than low arousal emotions. In summary, results suggest that the emotional

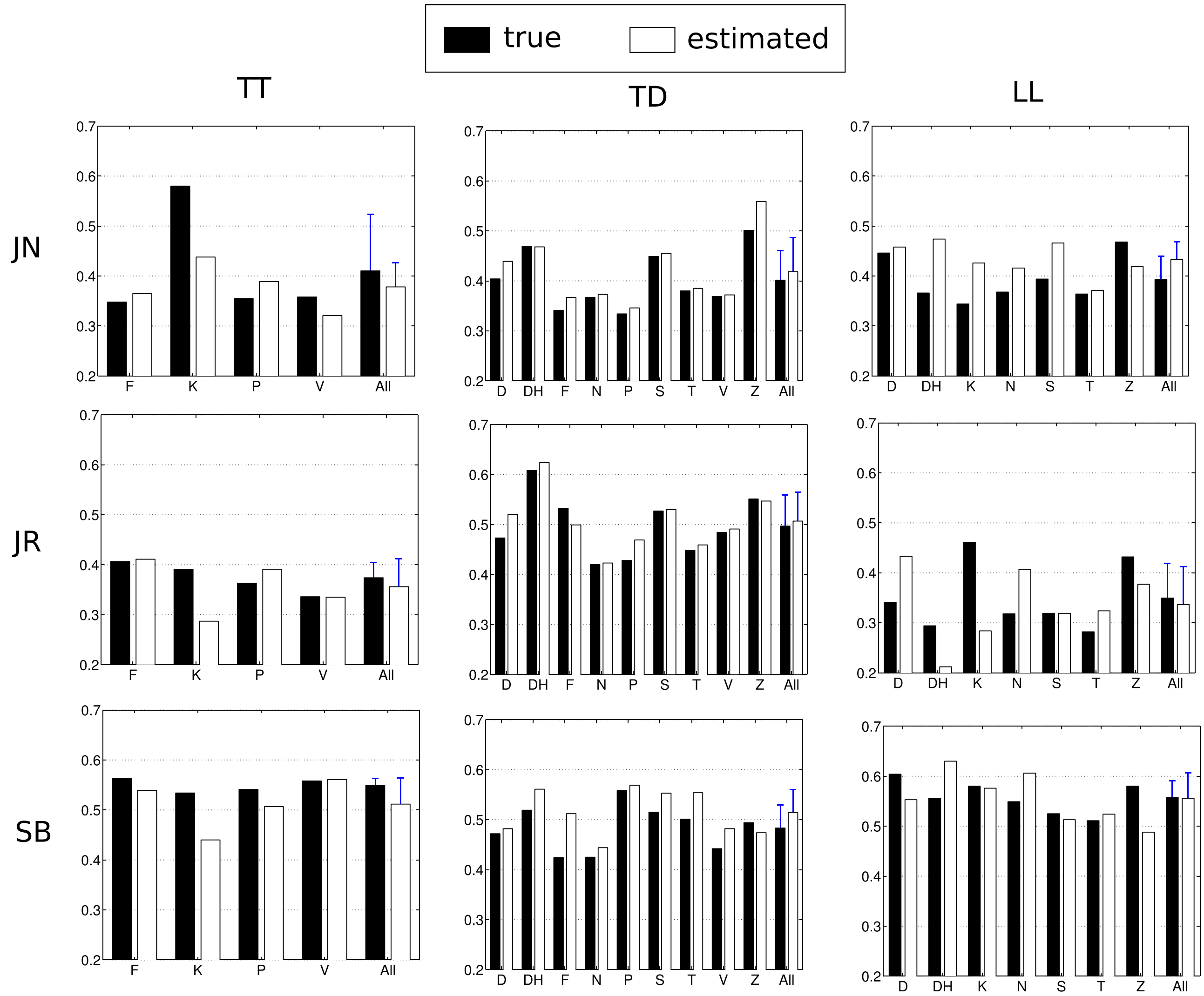

FIG. 14. (Color online) Unweighted emotion classification accuracy (%) for true and estimated data.

variation pattern of articulators depends on the linguistic criticality of the articulators.

Considering speaker-independent behaviors observed in this study, our experimental results also support the hypothesis that the emotional variation of articulatory positioning for vowels is associated with linguistic criticality of the tongue body (the tongue dorsum sensor parameters were used) in terms of the average of distance between the mean positions of different emotions. Supporting evidence was found by comparison between palatal and pharyngeal vowels in their movement directions (critical and non-critical) in terms of the average of centroid distances between emotion cluster pairs in Sec. V B. Larger variance of non-critical articulatory trajectories when compared with critical articulatory trajectories were reported in literature (Papcun *et al.*, 1992; Frankel and King, 2001), and this characteristic was employed for statistical identification of articulatory roles (Jackson and Singampalli, 2009). The experimental results in the present study provide additional information for vowels, i.e., the inter-emotion variance of articulatory positions is greater in the non-critical articulatory direction than in the critical articulatory direction. This suggests that the large variance of articulatory movements due to low linguistic criticality is an important factor of emotional modulation for vowels.

Previous studies have shown that tongue dorsum positions are dependent on emotion. For example, Erickson *et al.* (2000) reported that upward positioning and backward positioning of the tongue dorsum were observed for suspicion and admiration, respectively, in two vowels, /æ/ and /ʌ/ in an utterance "That's wonderful." The present study reports another speaker-independent characteristic in that the mean position of the tongue dorsum for velar stops is more forwarded and upward for hot anger than for other emotions. This tongue dorsum positioning for hot anger was consistently observed for all speakers only when the tongue dorsum were critical, implying that the exaggerated closure gesture of the tongue dorsum for velar stops is a characteristic of hot anger.

Non-critical articulators comprise dependent and redundant articulators according to the three-level categorization (critical, dependent, redundant) by Jackson and Singampalli (2009) and Guenther (1995). Dependent articulators refer to articulators whose movements are significantly dependent on

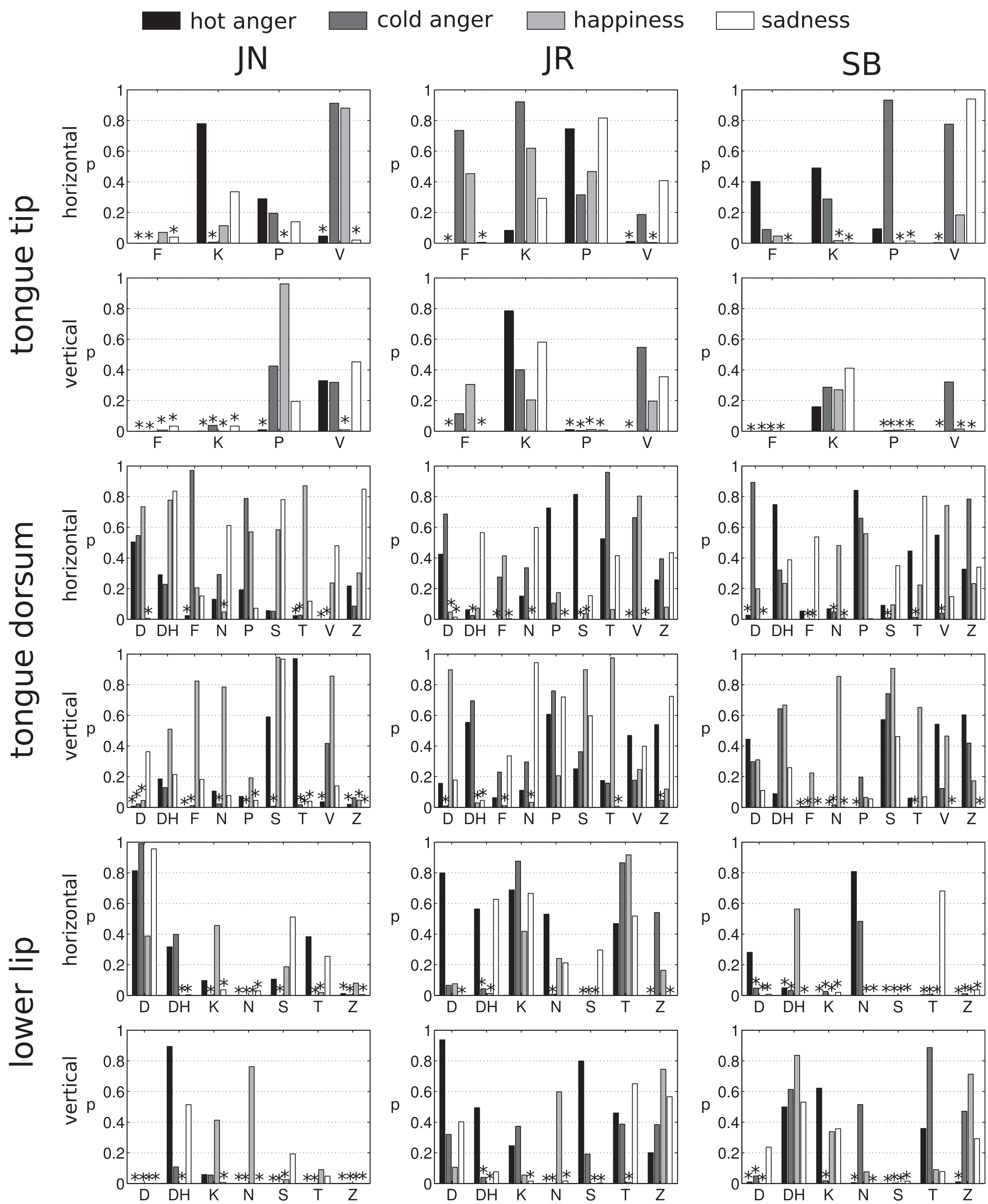

FIG. 15. t-statistic of the pair-sample t-test on two distributions, one of true data and the other of estimated data for non-critical articulators for each phone, each articulator, each emotion and each speaker. * indicates that p value is less than 0.05 for the case.

the movements of critical articulators due to anatomical structure and/or coordinated articulatory controls for linguistic encoding. For example, the tongue blade is a dependent articulator of the tongue tip, and the jaw is a dependent articulator of the lower lip in general. Redundant articulators refer to the remaining articulators whose movements are little dependent on the critical movements. Although the present study considered only critical and redundant articulators, controls of dependent articulators are also important to understand the detailed vocal tract shaping in emotional speech. Also this study did not consider the jaw, although previous studies (Erickson *et al.*, 2000; Erickson *et al.*, 2004; Erickson *et al.*, 2006) have shown that vertical jaw positioning is emotionally distinctive. In addition, jaw opening has been generally employed as a basic control of speech rhythm in literature (Nelson *et al.*, 1984; Fujimura and Erickson, 2004), so a better understanding of jaw movement can be useful for a comprehensive model that incorporates articulatory and rhythmic aspects of expressive speech. We have reported elsewhere (Kim *et al.*, 2014a) preliminary work on prosodic variations of emotional speech in the converter/distributor model framework (Fujimura, 2000), investigating temporal variations of articulatory movements depending on emotion, reflected in the model parameters.

The results of our analyses still cast an open question: What are the acoustic and perceptual consequences of the emotional variations of critical and non-critical articulators observed in the present study? Emotion perception tests with an articulatory synthesizer incorporating the controls of both critical and non-critical articulators will be useful for answering this question, although the articulatory synthesizer should be improved for minimizing potential loss of perceptual emotion quality first. To fully understand the variations of emotional speech production, it is important to know how the emotional variations of speech production components are related to each other, not only among articulators but also with other emotionally crucial voice cues, i.e., prosody (pitch, energy, and duration), intonation, and voice quality. Well known articulatory synthesizers, e.g., the ones developed by Rubin *et al.* (1996); Maeda (1982); Toutios and Narayanan (2013) do not yet incorporate para-linguistic aspects of expressive speech. These remain topics for future research. Also in the simulation experiment in Sec. VI, our model of estimating (linguistically) non-critical trajectories considers only the physiological constraints between critical articulators and *individual* non-critical articulator, although the physiological constraint among non-critical articulators, e.g., the tongue tip and the tongue dorsum when producing a labial consonant, also exists. Incorporating this constraint in the model is also our future work.

ACKNOWLEDGMENT

This work was supported by NSF Grant No. IIS-1116076.

APPENDIX: THE LIST OF SENTENCES OF OUR EMA DATABASE

The EMA database used in this study includes repetition of seven sentences. The list of the sentences is

(1) Say peep again? That's wonderful
(2) It was nine one five two eight nine five seven six two
(3) Say pop again? That's wonderful
(4) I saw nine tight night pipes in the sky last night
(5) Do not know how very joyful he was yesterday
(6) Say poop again? That's wonderful
(7) Native animals were often captured and taken to the zoo

The order of the seven sentences was randomized at each repetition of data collection.

Ananthakrishnan, G., and Engwall, O. (**2008**). "Important regions in the articulator trajectory," in *Proceedings of the International Seminar on Speech Production*, pp. 305-308.

Browman, C. P., and Goldstein, L. (**1989**). "Articulatory gestures as phonological units," Haskins Laboratories Status Report on Speech Research SR-99/100, Haskins Laboratories, New Haven, CT, pp. 69–101.

Busso, C., Lee, C., and Narayanan, S. S. (**2009**). "An analysis of emotionally salient aspects of fundamental frequency for emotion detection," IEEE Trans. Audio Speech Lang. Process. **17**(4), 582–596.

Erickson, D., Abramson, A., Maekawa, K., and Kaburagi, T. (**2000**). "Articulatory characteristics of emotional utterances in spoken English," in *Proceedings of the International Conference on Spoken Language Processing*, Vol. 2, pp. 365–368.

Erickson, D., Menezes, C., and Fujino, A. (**2004**). "Some articulatory measurements of real sadness," in *Proceedings of Interspeech*, Korea, pp. 1825–1828.

Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T., and Shibuya, Y. (**2006**). "Exploratory study of some acoustic and articulatory characteristics of sad speech," Phonetica **63**, 1–25.

Frankel, J., and King, S. (**2001**). "ASR—Articulatory speech recognition," in *Proceedings of Eurospeech*, Vol. 1, pp. 599–602.

Fujimura, O. (**2000**). "The C/D model and prosodic control of articulatory behavior," Phonetica **57**, 128–138.

Fujimura, O., and Erickson, D. (**2004**). "The C/D model for prosodic representation of expressive speech in English," in *Proceedings of Acoustical Society of Japan*, fall meeting, Okinawa, Japan, pp. 271–272.

Fujimura, O., Kiritani, S., and Ishida, H. (**1973**). "Computer controlled radiography for observation of movements of articulatory and other human organs," Comput. Biol. Med. **3**, 371–384.

Gobl, C., and Ni'Chasaide, A. (**2003**). "The role of voice quality in communicating emotion, mood and attitude," Speech Commun. **40**, 189–212.

Grimm, M., Kroschel, K., Mower, E., and Narayanan, S. (**2007**). "Primitives-based evaluation and estimation of emotions in speech," Speech Commun. **49**, 787–800.

Guenther, F. (**1995**). "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production," Psychol. Rev. **102**, 594–621.

Iskarous, K., Pouplier, M., Marin, S., and Harrington, J. (**2010**). "The interaction between prosodic boundaries and accent in the production of sibilants," in *Proceedings of the 5th International Conference on Speech Prosody*, Chicago, IL, pp. 1–4.

Jackson, P., and Singampalli, V. (**2008**). "Statistical identification of critical, dependent and redundant articulators," J. Acoust. Soc. Am. **123**, 3321.

Jackson, P. J., and Singampalli, V. D. (**2009**). "Statistical identification of articulation constraints in the production of speech," Speech Commun. **51**, 695–710.

Kim, J., Erickson, D., Lee, S., and Narayanan, S. (**2014a**). "A study of invariant properties and variation patterns in the converter/distributor model for emotional speech," in *Proceedings of Interspeech*, pp. 413–417.

Kim, J., Ghosh, P., Lee, S., and Narayanan, S. (**2012**). "A study of emotional information present in articulatory movements estimated using acoustic-to-articulatory inversion," in *Signal Information Processing Association Annual Summit and Conference*, Asia-Pacific, pp. 1–4.

Kim, J., Lee, S., and Narayanan, S. (**2014b**). "Estimation of the movement trajectories of non-crucial articulators based on the detection of crucial moments and physiological constraints," in *Proceedings of Interspeech*, pp. 163–168.

Kim, J., Lee, S., and Narayanan, S. S. (**2009**). "A detailed study of word-position effects on emotion expression in speech," in *Proceedings of Interspeech*, pp. 1987–1990.

Kim, J., Lee, S., and Narayanan, S. S. (**2010**). "A study of interplay between articulatory movement and prosodic characteristics in emotional speech production," in *Proceedings of Interspeech*, pp. 1173–1176.

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

Kruskal, W. H., and Wallis, W. A. (**1952**). "Use of ranks in one-criterion variance analysis," J. Am. Stat. Assoc. **47**, 583–621.

Lee, C. M., and Narayanan, S. S. (**2009**). "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process. **13**(2), 293–303.

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

Lee, S., Kato, T., and Narayanan, S. S. (**2008**). "Relation between geometry and kinematics of articulatory trajectory associated with emotional speech production," in *Proceedings of Interspeech*, Brisbane, Australia, pp. 2290–2293.

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

Maeda, S. (**1982**). "A digital simulation method of the vocal-tract system," Speech Commun. **1**, 199–229.

Nam, H., Goldstein, L., Saltzman, E., and Byrd, D. (**2004**). "Tada: An enhanced, portable task dynamics model in MATLAB," J. Acoust. Soc. Am. **115**(5), 2430.

Narayanan, S., Alwan, A., and Haker, K. (**1995**). "An articulatory study of fricative consonants using magnetic resonance imaging," J. Acoust. Soc. Am. **98**, 1325–1347.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (**2004**). "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**, 1771–1776.

Nelson, W., Perkell, J., and Westbury, J. (**1984**). "Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations," J. Acoust. Soc. Am. **75**, 945–951.

Papcun, G., Hochberg, J., Thomas, T., Laroche, F., Zacks, J., and Levy, S. (**1992**). "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," J. Acoust. Soc. Am. **92**(2), 688–700.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., and Jackson, M. (**1992**). "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078–3096.

Recasens, D., Pallars, M. D., and Fontdevila, J. (**1997**). "A model of lingual coarticulation based on articulatory constraints," J. Acoust. Soc. Am. **102**, 544–561.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., and Browman, C. (**1996**). "CASY and extensions to the task-dynamic model," in *First ESCA Tutorial and Research Workshop on Speech Production Modeling—Fourth Speech Production Seminar*, pp. 125–128.

Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust. Speech Signal Process. **26**, 43–49.

Saltzman, E., and Kelso, J. (**1987**). "Skilled actions: A task dynamic approach," Psychol. Rev. **94**, 84–106.

Saltzman, E., and Munhall, K. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecol. Psychol. **1**, 333–382.

Shami, M., and Verhelst, W. (**2007**). "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," Speech Commun. **49**, 201–212.

Stone, M. (**2005**). "A guide to analyzing tongue motion from ultrasound images," Clin. Ling. Phonet. **19**, 2128–2131.

Subtelny, J., and Oya, N. (**1972**). "Cineradiographic study of sibilants," Folia Phoniatr. **24**, 30–50.

Toutios, A., and Narayanan, S. (**2013**). "Articulatory synthesis of French connected speech from EMA data," in *Proceedings of Interspeech*, pp. 2738–2742.

Ververidis, D., and Kotropoulos, C. (**2006**). "Emotional speech recognition: Resources, features, and methods," Speech Commun. **48**, 1162–1181.

Williams, C. E., and Stevens, K. N. (**1972**). "Emotions and speech: Some acoustical correlates," J. Acoust. Soc. Am. **52**, 1238–1250.

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. S. (**2004**). "An acoustic study of emotions expressed in speech," in *Proceedings of Interspeech*, pp. 2193–2196.

Yuan, J., and Liberman, M. (**2008**). "Speaker identification on the scotus corpus," in *Proceedings of Acoustics*, pp. 5687–5690.