Proc APSIPA 2012
Hollywood, CA

A study of emotional information present in articulatory movements estimated using acoustic-to-articulatory inversion

Jangwon Kim* and Prasanta Ghosh† and Sungbok Lee* and Shrikanth S. Narayanan*
* Signal Analysis and Interpretation Lab. (SAIL), University of Southern California, Los Angeles, CA, U.S.A
E-mail: jangwon@usc.edu, Tel: +1-213-740-2318
† IBM Research India, Delhi, India
E-mail: prasantag@gmail.com

***Abstract*—This study examines emotion-specific information (ESI) in the articulatory movements estimated using acoustic-to-articulatory inversion on emotional speech. We study two main aspects: (1) the degree of similarity between the pair of estimated and original articulatory trajectories for the same and different emotions and (2) the amount of ESI present in the estimated trajectory. They are evaluated using mean squared error between the articulatory pair and by automated emotion classification. This study uses parallel acoustic and articulatory data in 5 elicited emotions spoken by 3 native American English speakers. We also test emotion classification performance using articulatory trajectories estimated from different acoustic feature sets and they turn out subject-dependent. Experimental results suggest that the ESI in the estimated trajectory, although smaller than that in the direct articulatory measurements, is found to be complementary to that in the prosodic features and hence, suggesting the usefulness of estimated articulatory data for emotions research.**

I. Introduction

Articulatory movements of vocal production carry a critical emotional information. For example, horizontal and/or vertical movement ranges (on the midsagittal plane), velocity and articulatory positions (e.g. tongue forwarding and/or lowering) vary depending on emotions [1], [2]. Also, the interplay between articulatory movements and voice source signal is useful for differentiating emotions, e.g. anger and happiness [3]. Nevertheless, emotion encoding in the articulation is not fully understood.

Despite the usefulness and potential of emotional speech, there are considerably fewer articulatory studies of emotional speech than those based on acoustic characteristics, presumably due to difficulties in obtaining direct articulatory data. There are several techniques for recording articulatory movements, for example ultrasound, electromagnetic articulography (EMA) [1] and real-time magnetic resonance imaging (rtMRI) [4], [5]. However, several practical issues remain for the acquisition of emotional speech and articulatory movement data. For example, the setup for EMA or rtMRI does not provide a natural speaking condition for a subject due to the instruments attached to subject's articulators or the unusual environment for data collection in an MRI scanner. This could contribute to increased challenges in eliciting emotional speech data.

Acoustic-to-articulatory inversion technique can be a promising venue to address these problems. The goal of acoustic-to-articulatory inversion is to estimate articulatory movement trajectories from acoustic speech signals and hence, it can be used to estimate articulatory details of emotional speech production alleviating the need for directly measuring articulatory movements. While there are several inversion algorithms available in the literature, e.g., [6], [7], [8], there has been no study to systematically understand how the inversion algorithms work with emotional speech and to what extent para-linguistic information, in particular emotional information, is preserved during inversion. Being automatic, acoustic-to-articulatory inversion can be used on a large corpus of emotional speech to obtain the corresponding articulatory movements. The potential benefits of successful inversion may also include the use of estimated articulatory parameters for emotion classification and, further, for better understanding of emotional speech production. Thus, in this work, we focus on studying the emotional information present in the articulatory movements estimated using acoustic-to-articulatory inversion.

For this purpose, in this study, we use our recently proposed acoustic-to-articulatory inversion method based on the generalized smoothness criterion (GSC) [6]. GSC works on the principle of constraining individual articulatory trajectories by the corresponding articulator-specific smoothness requirement. Since different articulators are smooth to different degrees [6], GSC provides an ideal framework to maintain articulator-specific smoothness in the estimated articulatory trajectories. Since articulatory movements are influenced by the type of emotion, the smoothness in articulatory movements may also get influenced depending on the emotional state of the subject. Using the GSC based inversion technique, we propose to examine to what extent GSC preserves the smoothness for different articulators under different emotions.

This study mainly aims at understanding the effectiveness of the inversion technique for emotional speech. We study two main aspects in this work: (1) How similar is the estimated trajectory to the original articulatory trajectory for various emotions? (2) How much does the estimated trajectory preserve emotion specific information (ESI)? Different mel frequency cepstral coefficient (MFCC) feature sets are tested to examine the influence of acoustic features used for the inversion of emotional speech. Experimental results suggest that the estimated articulatory trajectory carries ESI, although not as much as that in the original articulatory trajectory. This could be due to the fact that the emotional information in the articulatory trajectory estimated by inversion is upper-bounded by the emotional information in acoustic representation, and that there could be emotional information loss due to the inversion. In addition, when the estimated articulatory trajectory is used with some prosodic features, it generally improves the emotion classification accuracy indicating that the estimated articulatory movements provide ESI complementary to the prosodic features. We begin with describing the dataset used in this study.

II. Dataset

A. Parallel acoustic articulatory data

In the present study, we use articulatory data using electromagnetic articulography (EMA) from two female subjects (JN and JR) and a male subject (SB). All subjects are native speakers of American English and have had vocal training on acting which may be helpful for recording enough emotional speech within the limited time for data collection in practice. Seven sentences were prompted during the data collection. Each subject was asked to utter each individual sentence four to five times in each of three speaking styles, (normal, fast and loud) with each of 5 categorical emotions (neutrality (Neu), hot anger (HAng), cold anger (CAng), happiness (Hap), and sadness (Sad)). The subjects were asked to immerse themselves in a target emotion and speak the utterances when they were ready. The seven sentences of this dataset are following.

- Say peep again? That's wonderful.
- It was 9 1 5 2 8 9 5 7 6 2.
- Say pop again? That's wonderful.
- I saw 9 tight nightpipes in the sky last night.
- Don't know how very joyful he was yesterday.
- Say poop again? That's wonderful.
- Native animals were often captured and taken to the zoo.

The articulatory data used in this study consists of the position values from the horizontal (X) and vertical (Y) movement in the midsagittal plane of six lingual flesh-points (tongue tip, tongue blade, tongue dorsum, upper lip, lower lip and lower incisor), recorded at a rate of 200 Hz by EMA. Each sensor trajectory was smoothed using a 9th-order Butterworth filter with a 15 Hz cutoff frequency. Then, head movement and occlusal plane correction were applied to the sensor trajectory.

The emotion of each utterance audio is determined based on the evaluation results for the best representative emotion among the six categories, such as the list of five categorical emotions used by subjects and "others" (when none of the five emotions was the best representative emotion). This evaluation was done by four or five listeners who are also native speakers of American English. In this study, a native speaker of American English refers to those who were born in the United States and whose mother tongue and primary language are American English. The most representative emotion of each utterance was determined by majority voting of the evaluation results by five listeners. JN's data was collected following the same procedure of JR and SB as described in [9]. The more details regarding data collection, post-processing for head movement correction, occlusal plane correction, smoothing, and emotion evaluation results are described in [9].

The number of utterances for each emotion used in this study is given in Table 1.

TABLE I
The number of utterances for each of five categorical emotions, based on evaluation using majority voting [9].

Subject	Neu	HAng	CAng	Hap	Sad	Total
JN	66	59	101	76	78	380
JR	99	67	117	78	109	470
SB	95	80	69	67	87	398

B. Estimated articulatory data

The GSC based acoustic-to-articulatory inversion is used to estimate the articulatory trajectories for each utterance in the dataset. All utterances for each emotion of each subject in our database are divided into five folds using stratified sampling, i.e., each fold contains utterances from every emotion category in a balanced fashion. Then, articulatory trajectories for each utterance of one fold were estimated by the GSC based inversion algorithm[6]; the acoustic and original articulatory trajectories for utterances of the other four folds were used as the training data in the GSC. Thus for the entire corpus used in this study, we have the original as well as estimated articulatory trajectory for every utterance in each emotion spoken by all three subjects.

III. Hypotheses and Experimental Design

The goal of this study is to examine the quality of the articulatory trajectory estimated using acoustic-to-articulatory inversion for emotional speech. If the inversion algorithm is able to preserve the ESI, the estimated articulatory trajectory should also have ESI in it. Two main experiments are conducted to study the ESI present in the estimated trajectories: (1) Measuring similarity between original and estimated articulatory trajectories across different emotions and (2) Comparing the emotion classification accuracy using original articulatory trajectory and that using estimated articulatory trajectory.

For the first experiment, our hypothesis is that the articulatory trajectories (either original or estimated) corresponding to one sentence spoken several times with the same emotion are more similar to each other than those spoken with different emotions. This hypothesis was tested by comparing mean squared error (MSE) between two trajectories. The MSE (in mm) is computed as

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\sum_{j=1}^{J}(x_n^j - x_n^{j\star})^2} \quad (1)$$

where, $x_n^{j\star}$ and x_n^j are the j^{th} estimated and original articulatory trajectory values at n^{th} frame. N is the length of the utterance in number of frames. The number of different articulatory features (J) is 12, consisting of horizontal and vertical positions of 6 articulators. However, it is important to note that the durations of the utterances in different repetitions of the same sentence need not be identical; therefore, we need to align the utterances before computing MSE between two respective articulatory trajectories. This alignment is done at frame level using 12-dimensional MFCC (excluding the energy coefficient) feature and dynamic time warping (DTW) algorithm based on Mahalanobis distance measure. Utterance initial and the final silences are excluded for DTW. For a particular sentence spoken in different emotions, MSE is computed between the estimated and original trajectories (after alignment) across all emotions. If the two articulatory trajectories of the same emotion are more similar to each other than those of different emotions, then the lowest average MSE will be at the confusion cell of the same emotion.

The second experiment is to test how much ESI is maintained in the estimated articulatory trajectory compared to the original articulatory trajectory. This will indicate the effectiveness of the inversion algorithm in capturing ESI. The statistics, such as median, lower quartile, upper quartile and interquartile, of raw value and derivative of an articulatory trajectory are used as articulatory features. Gaussian mixture model (GMM) and Support Vector Machine (SVM) are used for emotion classification in a 5-fold cross-validation setup. We also investigated the amount of any additional ESI provided by the estimated articulatory trajectory in complement to that provided by the acoustic prosodic feature for emotion classification. Pitch, energy and MFCCs are extracted with a frame rate of 100 Hz (window size: 20 msec). Prosodic features consist of the statistics of energy, pitch and their derivatives. The same statistics used for articulatory trajectory are also used for pitch. The statistics for energy consist of mean, range, maximum and minimum. Quartiles are used for pitch and articulatory trajectory for minimizing the effects by their spurious noise. Apart from prosodic features, the same statistics used for energy are used for MFCCs (39-dimensional including Δ and $\Delta\Delta$ coefficients) as another acoustic feature set.

The articulatory, acoustic and prosodic features (statistics) are computed at the syllable level, word level, and utterance level. Phonetic boundary was estimated by HMM-based forced alignment using the P2FA tool first [10], and then they were manually corrected. Syllable segments were obtained using the NIST tsylb2 tool [11] which is a rule-based syllabification tool. Syllable, word and utterance segments are obtained from the segmented boundaries. Any region of pause and silence is excluded.

IV. Results and Discussion

A. Comparison of raw articulatory trajectories

The left (right) table in Table 2 shows the average MSE between two original (estimated and original) articulatory trajectories for each emotion pair in each subject's data. The articulatory trajectory estimated from 39 MFCCs is used for right table in Table 2. The results with other (12, 13 and 36 dimensional) MFCC sets show similar tendency. Based on the average difference between MSE of the same emotion and different emotions (avgDiff) and ANOVA analysis, we see that the average MSE of the pair of original articulatory trajectories of the same emotion is less than that of different emotions for most of the cases (p-value < 0.0005), which supports our hypothesis when only original trajectories are

TABLE II

The confusion matrices of average MSE between two articulatory trajectories of the same sentence by each subject. Two types of articulatory trajectory pairs are considered: (1) original-original, (2) estimated-original. The articulatory trajectory estimated from 39 MFCCs is used for "Estimated" in the right table as an example. The entry in the cell (i,j) of the matrices indicates the average MSE (averaged across all possible pairs of utterances for all sentences) between an original (or estimated) articulatory trajectory of i^{th} emotion from one fold and an original (or estimated) articulatory trajectory of j^{th} emotion from the remaining four folds. One-way ANOVA is used to test the difference between MSE for the same emotion pair and that for different emotion pair, which turns out to be significant at 99% level for most of the emotion pairs. F-measure and p-value are the results of ANOVA on MSE. Lowest MSE value in each row(column) of the matrices are indicated in bold font (underlined). avgDiff - average of (MSE of same emotion - MSE of diff emotion). Negative(or positive) avgDiff in a row/column indicates that on average MSE from the same emotion is lower(higher) than MSE from different emotion in that row/column.

				Original				
				Emo				
	Spkr			Neu	HAng	CAng	Hap	Sad
Original	JN	Emo	Neu	**2.26**	3.82	2.81	3.52	2.91
			HAng	3.82	**3.62**	3.69	3.88	3.85
			CAng	2.81	3.69	**2.72**	3.55	2.95
			Hap	3.52	3.88	3.55	**3.14**	3.50
			Sad	2.91	3.85	2.95	3.50	**2.90**
		avgDiff		-1.01	-0.19	-0.53	-0.47	-0.40
		F-measure		335	10	218	160	80
		p-value		0.000	0.001	0.000	0.000	0.000
	JR	Emo	Neu	**1.53**	2.19	2.15	2.60	2.15
			HAng	2.19	**2.17**	2.39	2.61	2.43
			CAng	**2.15**	2.39	2.34	2.84	2.55
			Hap	2.60	2.61	2.84	**2.40**	2.68
			Sad	2.15	2.43	2.55	2.68	**2.03**
		avgDiff		-0.74	-0.24	-0.14	-0.28	-0.42
		F-measure		1579	155	22	231	380
		p-value		0.000	0.000	0.000	0.000	0.000
	SB	Emo	Neu	**2.63**	4.26	3.44	3.45	2.82
			HAng	**4.26**	4.78	4.76	4.29	4.50
			CAng	3.44	4.76	3.50	4.36	**3.37**
			Hap	3.45	4.29	4.36	**3.15**	3.97
			Sad	2.82	4.50	3.37	3.97	**2.26**
		avgDiff		-0.86	0.33	-0.48	-0.87	-1.41
		F-measure		540	38	165	609	1564
		p-value		0.000	0.000	0.000	0.000	0.000

				Original							
				Emo							
	Spkr			Neu	HAng	CAng	Hap	Sad	avgDiff	F-measure	p-value
Estimated	JN	Emo	Neu	**2.46**	3.99	2.93	3.36	2.66	-0.78	789	0.000
			HAng	**3.00**	3.77	3.27	3.41	3.01	0.60	824	0.000
			CAng	**2.64**	3.92	2.94	3.36	2.68	-0.21	39	0.000
			Hap	3.04	4.02	3.36	3.41	**3.01**	0.05	7	0.008
			Sad	2.86	4.04	3.12	3.42	**2.72**	-0.64	493	0.000
		avgDiff		-0.43	-0.22	-0.24	0.03	-0.12			
		F-measure		306	67	136	4	16			
		p-value		0.000	0.000	0.000	0.046	0.000			
	JR	Emo	Neu	**1.96**	2.54	2.53	2.83	2.30	-0.59	1285	0.000
			HAng	**2.08**	2.45	2.52	2.63	2.23	0.09	34	0.000
			CAng	**2.10**	2.57	2.54	2.81	2.31	0.09	35	0.000
			Hap	**2.41**	2.71	2.77	2.70	2.47	0.11	52	0.000
			Sad	2.25	2.70	2.74	2.81	**2.22**	-0.41	564	0.000
		avgDiff		-0.25	-0.18	-0.10	-0.07	-0.11			
		F-measure		537	230	11	14	49			
		p-value		0.000	0.000	0.000	0.000	0.000			
	SB	Emo	Neu	2.77	4.15	3.44	3.56	**2.64**	-0.68	668	0.000
			HAng	3.19	4.21	3.81	3.64	**3.12**	0.77	1049	0.000
			CAng	2.95	4.28	3.42	3.75	**2.68**	0.01	0.25	0.612
			Hap	**2.96**	4.04	3.65	3.34	3.02	-0.08	5.95	0.015
			Sad	2.93	4.28	3.48	3.84	**2.53**	-1.10	1848	0.000
		avgDiff		-0.24	0.02	-0.18	-0.36	-0.34			
		F-measure		239	2.66	78	348	364			
		p-value		0.000	0.103	0.000	0.000	0.000			

considered. However, there are some exceptions. For example, the lowest MSE for cold anger is with sadness (3.37 in SB) or neutrality (2.15 in JR), presumably due to their similar nature of articulatory movements in arousal dimension: both of them lie in the low arousal dimension and neutrality overlaps with other emotions.

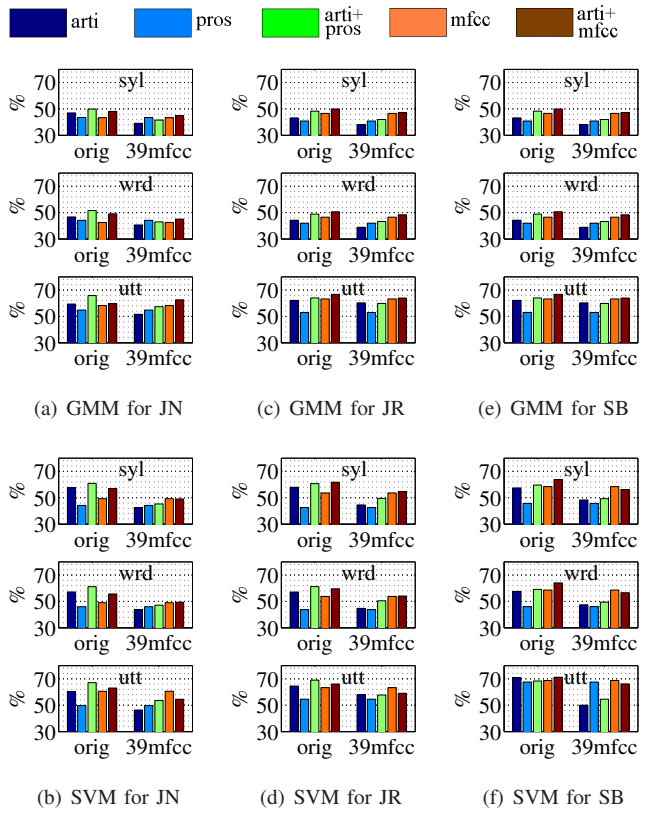
The MSE confusion matrix for estimated-original case is not symmetric and hence we perform ANOVA analysis for each row and each column unlike the original-original case. Most of the lowest average MSE values in the row/column of MSE matrices (for JN, JR, and SB) for estimated-original case do not occur in the same emotion cell. This is also reflected in many of the avgDiff values being positive unlike in the case of original-original. Thus the avgDiff values and ANOVA analysis in the right table in Table 2 indicate that the emotional contrast in terms of average MSE for estimated-original articulatory pair case (right table) is less than that for the two original articulatory pair case (left table). This could be attributed to the approximations and limitation inherent in the inversion.

Among five cells of each row of low arousal emotions (cold anger and sadness) in the right table in Table 2, the smallest MSE is at the cell of low arousal emotions or neutrality. However, among five cells of each row of high arousal emotion (hot anger and happiness), the smallest MSE is not at the cell of high arousal emotions, rather at the cell of low arousal emotion or neutrality. This result indicates that the similarity of estimated articulatory trajectories of all emotions are greater with original articulatory trajectories of low arousal emotions than those of high arousal emotions. It could be because the inversion does not well maintain the nature of emotional contrast of original articulatory trajectory in the arousal dimension. 'avgDiff' values corresponding to the columns of MSE matrices (in right table) are mostly negative while those corresponding to the rows are not. It indicates that an estimated articulatory trajectory is not always close to the original one of the same emotion in Euclidean distance measure (MSE) but an original articulatory trajectory is most similar to the estimated articulatory trajectory of the same emotion. Since the 'avgDiff' across columns and rows in estimated-original case show different characteristics, we further investigate the ESI in estimated trajectories by emotion classification experiment.

B. Emotion classification

Figure 1 shows emotion classification accuracies using features derived from statistics of articulatory, prosodic, and MFCC trajectories as well as joint articulatory-prosodic and articulatory-MFCC trajectories. The feature dimension in each of these cases varies, hence, we used principal component analysis to select a reduced set of features which explain 90% of the feature variance. We use both generative (GMM) and discriminative (SVM) classifiers for the emotion classification task to investigate the benefit of two distinct types of classifiers. From Figure 1, we see that the relative classification accuracies across different feature sets are similar for both GMM and SVM but in some cases SVM achieves higher classification accuracies compared to that of GMM (e.g., JN syl, JR syl, JR wrd level).

Among two acoustic-only features used for emotion classification, accuracies using 'mfcc' is higher than that using 'prosodic' features. This indicates that 'mfcc' carries richer ESI compared to 'prosodic', although the later is often used for emotion classification in the literature. On the other hand, for features involving articulatory representations (i.e., 'arti', 'arti+pros', 'arti+mfcc'), we see that the accuracies using original articulatory data ('orig') is more than that using estimated ones ('39mfcc'). This indicates that the estimated articulatory data have ESI but not to the same degree as in original articulatory data. From Figure 1, we also see that accuracies increase by using 'arti+pros' compared to 'pros' (except in 'utt' case for SB) when estimated articulatory data is used ('39mfcc'). However, similar improvement in 'arti+mfcc' compared to 'mfcc' occurs only in the case of GMM classifier. This implies that the estimated articulatory data has ESI complementary to the prosodic features, but not always true for the acoustic features. Considering the results using GMM classifiers, it is interesting since the articulatory features are estimated from the acoustic features using inversion. This could be due to the nonlinear map between acoustic and articulatory spaces and inversion



(a) GMM for JN (c) GMM for JR (e) GMM for SB

(b) SVM for JN (d) SVM for JR (f) SVM for SB

Fig. 1. *Emotion classification accuracies (in percentage) for three subjects with various combinations of acoustic (mfcc), prosodic (pros) and articulatory (arti) features computed at syllable level (syl), word level (wrd), and utterance level (utt). 'orig' corresponds to the case when articulatory features are computed from the original articulatory trajectories. '39mfcc' corresponds to the case when we use articulatory trajectories estimated from 39-dimensional MFCC+Δ+$\Delta\Delta$. We also experimented with 12-dim, 13-dim, and 36-dim MFCC and their derivatives but the accuracy patterns did not change much. Accuracies using 'mfcc' and 'pros' are identical in both 'orig' and '39mfcc' cases; accuracies change only when 'arti' is considered. Accuracies are reported using GMM and SVM as classifiers for each subject.*

could preserve ESI in the estimated articulators which is not in the acoustic data.

It is also interesting to note that most of the times the accuracies using articulatory features for 'orig' case is greater than those using acoustic and prosodic features. So articulatory features have more discriminatory power for emotion classification compared to acoustic prosodic features. But this is not the case when estimated articulators ('39mfcc') are considered. Thus, the ESI in estimated articulatory features is lower than but complementary to that in acoustic prosodic features.

From the percentage difference in classification accuracy between 'orig' and '39mfcc' (e.g., for SVM in 'utt' level, SB: 20.83%, JN: 14.01%, JR: 6.63%), we speculate that the "quality" of the estimated trajectories appears to be subject-dependent. The MFCCs showed greater classification accuracy for SB than the other subjects (SB: 70.27%, JN: 58.21%, JR: 63.26%). Although MFCCs for SB carry highest amount of ESI, articulatory data derived from MFCCs resulted in maximal drop in classification performance, indicating that the inversion was less successful in capturing ESI for SB's data than other subjects' data.

The benefit of energy terms in MFCCs for inversion is also subject-dependent. For SB, the articulatory trajectories estimated with 12-dim MFCCs and 36-dim MFCCs show better classification accuracy than those estimated with 13-dim MFCCs and 39-dim MFCCs respectively in all levels. But this is not the case for JN and JR (e.g. the 'wrd' level accuracies using articulatory features estimated from 12-dim MFCCs, 13-dim MFCCs, 36-dim MFCCs and 39-dim MFCCs by SVM are following: 48.28%, 46.76%, 48.47%, 47.49% for SB, 41.93%, 44.66%, 41.71%, 43.89% for JN, 44.48%, 43.96%, 44.63%, 44.76% for JR).

Classification accuracy using utterance level features is the highest among the considered syllable, word and utterance level features. Word level features is slightly better than syllable level features in general. It indicates that utterance level statistics have the maximal information of emotional contrast in original as well as estimated articulatory features.

V. Conclusions and Future Works

From this study, we found that the articulatory movements estimated using GSC based inversion carry important emotion specific information (ESI), but it is smaller when compared to the original articulatory movements. Emotion classification showed that the ESI in the estimated articulators offers complementary emotion information to that in the prosodic features. This is encouraging since estimated articulators can be used for emotion study in cases when direct (original) articulatory measurements are not available. However, further investigations are required to fully understand the pros and cons of using estimated articulatory data for emotion research, particularly discerning how ESI is encoded in estimated articulatory movements e.g., is it in dimensional way or categorical way of emotion? We would also like to to investigate other inversion techniques for studying ESI in the estimated articulators because the present study showed there is room for improving emotion classification performance using estimated articulators. In this study we have used subject-dependent inversion technique; a rich research direction would be to investigate ESI in estimated articulators when subject-independent inversion [12] is used. This is because in subject-independent inversion, we do not need training data for inversion from the subject, for whom we need to estimate articulatory movements, which is often useful in practice.

Acknowledgments

This work is supported by the NSF and the NIH.

References

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

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

[3] Kim, J., Lee, S., Narayanan, S. S., "A study of interplay between articulatory movement and prosodic characteristics in emotional speech production," *in Proceedings of Interspeech*, pages 1173-1176, 2010.

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

[5] Lee, S., Bresch, E., Adams, J., Kazemzadeh, A., and Narayanan, S. S., "A study of emotional speech articulation using a fast magnetic resonance imaging technique," *in Proceedings of InterSpeech*, 2005

[6] Ghosh, P. K., Narayanan, S. S., "A generalized smoothness criterion for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2162–2172, 2010.

[7] Toda, T., Black, A., Tokuda, K., "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.* 50, 215-217, 2008.

[8] Yehia, H., "A study on the speech acoustic-to-articulatory mapping using morphological constraints," Ph.D. thesis, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan, 2002.

[9] Kim, J., Lee, S., Narayanan, S. S., "An exploratory study of the relations between perceived emotion strength and articulatory kinematics," *in Proceedings of Interspeech*, Florence, Italy, 2011.

[10] Yuan, J., Liberman, M., "Speaker identification on the SCOTUS corpus", in Proceedings of Acoustics," 08, pages 5687-5690, 2008.

[11] Fisher, W., "A C implementation of Daniel Kahns theory of English syllable structure," http://www.itl.nist.gov/iad/894.01/tools/

[12] Ghosh, P. K., Narayanan, S. S., "A subject-independent acoustic-to-articulatory inversion", *in Proceedings ICASSP*, Prague, Czech Republic, 22-27 May, 2011.