Speaker verification based on fusion of acoustic and articulatory information

Ming Li[1]*, Jangwon Kim*[1]*, Prasanta Ghosh*[2]*, Vikram Ramanarayanan*[1] *and Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA
[2]Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India
{mingli,jangwon,vramanar}@usc.edu, prasantg@ee.iisc.ernet.in, shri@sipi.usc.edu

Abstract

We propose a *practical*, feature-level fusion approach for combining acoustic and articulatory information in speaker verification task. We find that concatenating articulation features obtained from the measured speech production data with conventional Mel-frequency cepstral coefficients (MFCCs) improves the overall speaker verification performance. However, since access to the measured articulatory data is impractical for real world speaker verification applications, we also experiment with estimated articulatory features obtained using acoustic-to-articulatory inversion technique. Specifically, we show that augmenting MFCCs with articulatory features obtained from subject-independent acoustic-to-articulatory inversion technique also significantly enhances the speaker verification performance. This performance boost could be due to the information about inter-speaker variation present in the estimated articulatory features, especially at the mean and variance level. Experimental results on the Wisconsin X-Ray Microbeam database show that the proposed acoustic-estimated-articulatory fusion approach significantly outperforms the traditional acoustic-only baseline, providing up to 10% relative reduction in Equal Error Rate (EER). We further show that we can achieve an additional 5% relative reduction in EER after score-level fusion.

Index Terms: speech production, speaker verification, articulation features, acoustic-to-articulatory inversion, biometrics

1. Introduction

The goal in a speaker verification (SRE) task is to determine whether a given segment of speech is spoken by the claimed target speaker.

At the acoustics level, joint factor analysis (JFA) [1, 2, 3] has contributed to the state-of-the-art performance in the text independent SRE. It is a powerful and widely used technique for compensating the acoustic variabilities caused by different channels and sessions. Recently, total variability i-vector modeling has gained significant attention in SRE due to its excellent performance, low complexity and small model size [4]. In this approach, a single factor analysis is used as a front-end to generate a low dimensional total variability space (i.e. the i-vector space) which jointly models speaker and channel variabilities [4]. The factor analysis can also be extended to a simplified supervised version to enhance the performance and reduce the computational cost [5]. Within the i-vector space, variability compensation methods, such as Within-Class Covariance Normalization (WCCN) [6] and Linear Discriminative Analysis (LDA), are performed to reduce the variability for the subsequent probabilistic LDA (PLDA) modeling [7, 8]. Sparse representation could also be applied in the SRE task [9, 10, 11].

In addition to the aforementioned state-of-the-art modeling methods, various kinds of features have also been proposed for text independent speaker recognition (e.g. short-term spectral features, voice source features, spectral-temporal features, prosodic features and high-level features) [12]. Both feature level and score level fusion based on multiple features have been shown to enhance the overall SRE system performance [12]. In this work, our goal is to examine the utility of speech production-oriented features for the SRE task.

Several studies have shown that an important source of inter-speaker variability in speech acoustics lies in the variability in the vocal tract morphology across various speakers; morphological variability could result from the differences in the vocal tract length [13, 14, 15, 16], the morphology of the hard palate and the posterior pharyngeal wall [17, 18, 19]). Since vocal tract length is closely related to the formant frequency [16, 14], change in vocal tract length scales the frequency of the speech spectra for voiced sounds. This has been extensively used for vocal tract length normalization (VTLN) [20, 21] in automatic speech recognition (ASR). Unlike normalization, we focus on exploiting morphological variations as the cue for speaker characteristics in SRE applications in this paper. Speakers with flat palates exhibit less articulatory variability during vowel production than speakers with highly domed palates [22, 23, 24, 25]. Articulation of coronal fricatives is also influenced by palate shape, including apical vs. laminal articulation of sibilants [26], as well as jaw height and the positioning of the tongue body [27, 28]. Therefore due to the different vocal tract morphology characteristics, different speakers articulate even the same words differently. We also show in [29] that the fusion of speech and articulation features enhances discrimination between different morphological structures. This motivates us to examine the utility of morphological characteristics in the SRE task by using articulatory features in addition to acoustics.

We find that concatenating articulation features obtained from the measured speech production data with conventional Mel-frequency cepstral coefficients (MFCCs) improves the overall SRE performance. However, since measuring articulatory movement during speech production is impractical for real world SRE applications, SRE experiment is also performed where the measured articulatory features are replaced with estimated articulatory features obtained using acoustic-to-articulatory inversion. Specifically, we show that augmenting MFCCs with features obtained from subject-independent acoustic-to-articulatory inversion techniques significantly enhances the SRE performance. In other words, we show that the estimated articulation obtained by articulatory inversion carry useful information about inter-speaker variation, especially at the mean and variance level, which leads to better performance.

This work was supported in part by NIH Grant DC007124, NSF and Department of Justice.

To our best knowledge, there has not been a study on this topic.

Although the estimated articulatory features are also generated from speech signals, we can show that adding this new information (articulation-acoustics mapping) on top of MFCCs can still enhance the performance. Theoretical supports from machine learning fields are provided in [30, 31] which show that the recognition of target label can be improved with additional knowledge of related labels. Practically, this concatenation based speech-articulatory feature level fusion has been reported to increase the ASR performance [32, 33] significantly.

Next section describes the data set partition and experimental setup of the adopted X_Ray Microbeam database. Feature extraction, subject-independent articulatory inversion and modeling methods are described in Section 3. Section 4 presents the experimental results and discussion. Some concluding remarks are given in Section 5.

2. Data

We used the Wisconsin X-Ray Microbeam data (XRMB) [34] for our analysis and experiments. A key feature of this database is that both articulatory measurement and simultaneously recorded speech signal are available from multiple speakers. We selected read speech data (citation words, sentences and paragraphs) from sessions 1 to 101 for each speaker from JW11 to JW63 which resulting in a total of 4034 utterances from 46 speakers with an average duration of 5.72 seconds per utterance. Note that we excluded speech sessions involving different speaking styles (such as fast or slow speech, emphasized speech, or stimuli that involved diadokinesis). We also omitted speaker sessions where a speaker had to repeat an utterance, as well as those which were found to contain severe pellet tracking errors, as detailed in the XRMB Manual [34].

Table 1 shows the two protocols that we adopted in the evaluation, namely "ALL" and "L5S". We used all sessions from speaker JW11 to JW40 (26 speakers with 2295 utterances) as the background data and select the session 11 (a paragraph session) of each speaker from JW41 to JW63 (20 speakers) as the target registration utterance. For the testing, protocol "ALL" selects all the sessions (excluding session 11 in the target set) from speaker JW41 to JW63 (a total of 20 speakers and 1719 utterances) while protocol "L5S" only adopts those sessions that are more than 5 seconds long (a total of 20 speakers and 840 utterances). The reason to create a separate "L5S" protocol is that some testing utterances in the "ALL" protocol are too short. In order to perform Test Segment Score Normalization (T-norm), we selected all other paragraph sentences (session 12,79,80,81, totally 95 utterances) in the background set as the T-norm set.

To evaluate the performance of the acoustic-only baseline as well as the acoustic-estimated-articulatory system, we followed the protocol (as shown in Table 1) exactly. For the speech-real-articulation system, a subset of data were removed from the train, target and testing sets due to missing data in some articulatory channels [34]. We name this modified "ALL" protocol as "ALL-small" protocol. In "ALL-small" protocol, each utterance is shorter and there are 1849, 18, and 1389 utterances in train, target and testing sets, respectively.

Table 1: *Data set partition for SRE experiments. Other L5S sessions of JW41-63 denotes all the longer than 5 seconds sessions (exclude 11) of speaker JW41-JW63.*

Data sets & Protocol	ALL	L5S
Background: all sessions of JW11-40	√	√
Target: session 11 of JW41-63	√	√
Test: other sessions of JW41-63	√	
Test: other L5S sessions of JW41-63		√
Tnorm: sessions 11,12,79,80,81 of JW11-40	√	√

3. Experimental Setup

3.1. Subject-independent inversion

We used the generalized smoothness criterion (GSC) for acoustic-to-articulatory inversion [35] under a subject-indepedent inversion setting [36]. The GSC estimates articulatory parameters given acoustic features so that the estimated parameters are optimal solution which satisfies two conditions: (1) the estimated trajectories are smooth and slowly varying and (2) the difference between the estimated and original articulatory parameters weighted by similarity metric of corresponding acoustic features is minimum. The subject-independent inversion setting uses a probability feature vector (PFV) for acoustic features. PFV is a normalized likelihood score of the conventional acoustic feature vector, i.e. MFCCs, to the 40 clusters of a general acoustic model (GAM) [36]. The general acoustic model represents the variabilities in acoustic space, which was created with TIMIT data [37].

In subject-independent acoustic-to-articulatory inversion, the acoustic of an arbitrary test subject is converted to a PFV which is then used to find the closest PFV from the chosen exemplar whose articulatory data is used for training the inversion mapping. It is expected that the PFV reflects the acoustic sound produced by the test subject irrespective of the speaker, i.e., the PFVs corresponding to a sound recorded from different speakers including the exemplar should be similar to each other so that the speaker variability is eliminated in the inversion. The quality of this speaker variability elimination solely depends on the generalizations of the GAM used to compute the probability feature vector. Note that the GAM used in this work is built using the TIMIT training corpus whereas the articulatory inversion is done on XRMB corpus which may have a different acoustics than that of TIMIT resulting in a poor elimination of speaker variability during inversion. This in turn gets reflected in the estimated articulatory features which when used for SRE task provides inter-speaker discrimination in addition to MFCC. Thus the SRE performance improvement (Sec4) in this work may results from the non-linear mapping between acoustic and articulatory spaces and the residual speaker specific information present in the probability features computed during the subject-independent acoustic-to-articulatory inversion.

3.2. Articulatory features

We used tract variables for articulatory parameters as in previous study [36]. The tract variables are estimated from an Electromagnetic Articulography database [38], which includes speech audio spoken by a native female speaker of American English and its parallel articulatory data. The speaker was asked to read 460 English sentences (approximately 69 minutes) identical to the sentences of MOCHA TIMIT database [39]. The tract variables include nine articulatory parameters, such as lip aperture (LA), lip protrusion (PRO), jaw opening (JAW_OPEN), the constriction degree (CD) and constriction location (CL) of tongue tip (TT), tongue blade (TB), and tongue dorsum (TD). The constriction location parameter for each tongue sensor is

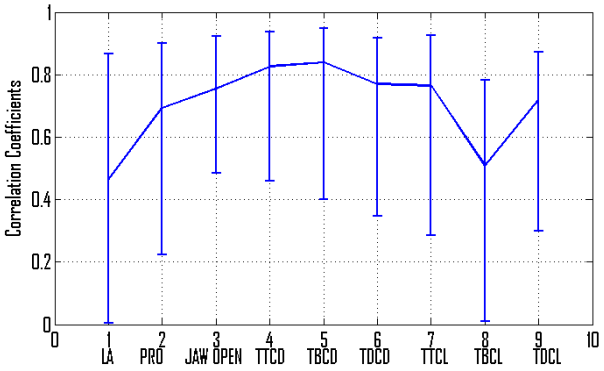

Figure 1: *Errorbar of pair-wise correlation coefficients between session one estimated articulatory signals (after DTW) from all 46 speakers (all speak the same word sequence, totally 1081 pair-wise DTW and correlation). The nine dimensions of the estimated articulation are LA, PRO, JAW_OPEN, TTCD, TBCD, TDCD, TTCL, TBCL, and TDCL, respectively.*

the distance from a fixed point on the palatal line, which is manually chosen by visual inspection, to the projected point of each sample to the palatal line. We followed the definitions in the previous study [36] for the other parameters, such as LA, PRO, JAW_OPEN, CDs.

Fig. 1 shows the error bar plot of pair-wise correlation coefficients between the estimated articulatory signals of the session one (the data of 46 speakers) after temporal alignment on the utterance pairs. All spoke the same word sequence in this case, allowing us to compare the inter-speaker variations by this method. Dynamic time warping (DTW) (applied on the estimated articulation) was used to remove possible speaking-rate confounds for this correlation study. Fig.1 shows that tongue constriction location features (dim 7,8,9) have more inter-speaker variations than tongue constriction degree features (dim 4,5,6). Lip aperture and tongue body constriction location (dim 1,8) show relatively less correlation, implying that their inter-speaker variations are larger than the other tract variables.

Fig. 2 shows the estimated articulatory signals after DTW on LA and TBCL of session one from the two-speaker pairs. Within all the speaker pairs, the pair of speaker JW48 and 33 has the highest correlation, while the pair of speaker JW 48 and JW 59 has the lowest one (JW48, 33 are female, JW59 is male). We can see from Figs. 2 (a)-(b) that even with the highest correlation, their estimated articulatory signals are not exactly the same. We investigated this further and found that the mean and variance values of these signals could differentiate between different speakers to a large extent, as shown in Figs. 2 (c)-(d).

In order to test our assumption that the mean and variance might carry the information of inter-speaker variability, we performed a simple multi-class SVM experiment. Table 2 shows the performance of speaker classification with different utterance-level features derived from estimated articulatory trajectories. The number of speaker classes is 26. Sessions 12, 79, 80 and 81 of all 26 speakers in the background data set (this is actually the T-norm set) was used for train set, and session 11 were used for test data. Table2 shows the performance of 3 systems based on different utterance-level features. By using only mean and variance, system 2 achieve around 50% accuracy, indicating that they do carry valuable information regarding inter-speaker variations. This result may also suggest to normalize mean and variance of estimated articulatory parameters for minimizing speaker-dependent information.

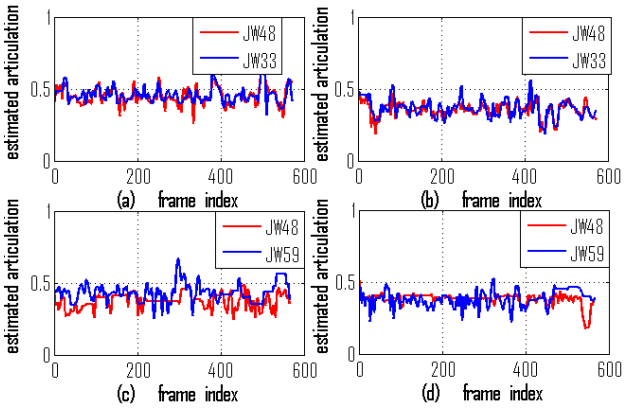

Figure 2: *Estimated articulatory signals (after DTW) of lip aperture and tongue body constriction location of session one from two-speaker pairs. The top plots are for the pair of speaker JW 48 and 33, and the bottom plots are for the pair of speaker JW 48 and 59. The first speaker pair shows high correlation, while the other pair shows low correlation. (a) LA of file TP001_2 from JW48 and 33; (b) TBCL of file TP001_2 from JW48 and 33; (c) LA of file TP001 from JW48 and 59; (d) TBCL of file TP001 from JW48 and 59.*

Table 2: *The performance of 26 speaker classes (closed set) identification systems based on different utterance-level features derived from estimated articulatory data.*

Features & Systems	1	2	3
mean	√	√	√
variance		√	√
mean crossing rate			√
Accuracy	32%	48%	52%

3.3. Front end processing

Wiener filtering [40] was adopted for X-Ray Microbeam data to reduce stationary artifact noises. After voice activity detection (VAD), non-speech frames were eliminated and cepstral features were extracted. Real and estimated articulatory signals were also truncated based on the VAD results and then resampled at 100 hz. A 25ms Hamming window with 10ms shifts was adopted for MFCC extraction. Each utterance was converted into a sequence of 36-dimensional feature vectors, each consisting of 18 MFCC coefficients and their first derivatives. Cepstral mean subtraction and variance (MVN) normalization were performed to normalize the MFCC and real articulatory features to zero mean and unit variance on a per utterance basis. The reason to perform MVN on the real articulatory signals is that the baseline values of these sensors have already encoded the vocal tract shape information of the speakers' [34]. For the estimated articulatory signals, we do not perform MVN since they are calculated from speech signals and mean variance are useful for speaker recognition. After MVN, MFCCs are con-

Table 3: *Performance of MFCC-real-articulation system with "ALL-small" protocol*

ID	Systems	"ALL-small" (%)	
		OptDCF	EER
1	MFCC-only	11.04	11.95
2	MFCC-real-articulation	9.98	10.15
3	Score level fusion 1+2	6.42	6.77

catenated with real or estimated articulatory signals to generate the MFCC-real-articulation and MFCC-estimated-articulation these two enhanced feature sets.

3.4. GMM baseline modeling

A UBM in conjunction with a MAP model adaptation approach [14] was used to model different speakers in a supervised manner. All the data in the background set was adopted to train a 256-component UBM, and MAP adaptation was performed using the training set data for each speaker. A relevance factor of 16 was used for the MAP adaptation. We performed AT-norm to calibrate the scores. Every testing utterance is scored on every target sample to generate the trials. The reason to use the GMM baseline here rather than the state-of-the-art I-vector PLDA method is that the data set is too small to train a large scale factor analysis model.

4. Experimental Results and Discussions

We evaluate the system performance in terms of identification weighted accuracy, verification EER and old OptDCF cost value [41]. Table3 shows the performance on the MFCC only as well as the MFCC-real-articulation features systems with the "ALL-small" protocol. We can see that by augmenting with the measured articulatory features (although mean and variance normalized), the enhanced feature set reduced the EER from 11.95% to 10.15%. Score level fusion of these two systems further reduced the EER to 6.77%, a 40% relative EER reduction compared to the MFCC only system. Thus it is clear that, adding real articulation information enhances the SRE performance.

Table 4 and 5 show the SRE performance when estimated articulatory features are used in augmenting MFCCs in "ALL" and "L5S" protocols respectively. Here we observe similar patterns as in Table3. The MFCC-estimated-articulation features also achieved 4% and 8% relative EER reduction for "ALL" and "L5S" protocols, respectively. Score level fusion further increased this relative reduction to 9% and 14%. However, it should be noted that the improvement is not as big as the real articulation case in Table3. The SRE performance improvement using MFCC-estimated-articulation, though moderate, suggests the potential benefit that estimated articulatory features may provide in SRE task. This is particularly important because in a real world SRE application, we have only access to the speech signal. In such scenarios it is only articulatory inversion that can provide information about speaker's morphological characteristics in terms of the estimated articulatory features. And experimental results in this work show that estimated articulatory features indeed provide production oriented information (complementary to the MFCCs) to discriminate different speakers. This is also shown by the Detection Error Trade-off (DET) curves in Fig.3 , which clearly demonstrates that adding estimated articulatory features improves the SRE performance.

Table 4: *Performance of MFCC-estimated-articulation system with "ALL" protocol*

ID	Systems	"ALL" (%)		
		OptDCF	EER	Accuracy
1	MFCC-only	8.68	8.73	89.65
2	MFCC-estimated-articulation	8.40	8.44	90.92
3	Score level fusion 1+2	7.83	7.91	91.74

Table 5: *Performance of MFCC-estimated-articulation system with "L5S" protocol*

ID	Systems	"L5S" (%)		
		OptDCF	EER	Accuracy
1	MFCC-only	4.84	4.88	95.95
2	MFCC-estimated-articulation	4.34	4.52	97.14
3	Score level fusion 1+2	4.05	4.17	97.02

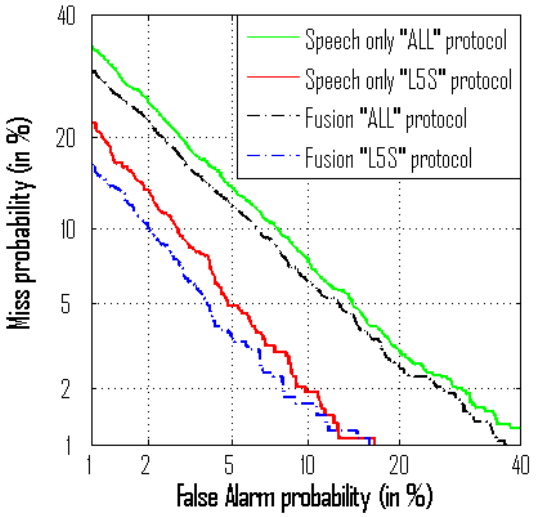

Figure 3: *DET curves of speech only (ID1) and fusion (ID3) results in Table 4 and 5.*

5. Conclusion

We propose a practical feature-level fusion approach for speaker verification using information from both acoustic and articulatory representations. We find that the speaker verification performance improves by concatenating articulation features from measured articulatory movement data during speech production with conventional MFCCs. However, since access to the measured articulatory movement is impractical for real world speaker verification applications, we also experiment with estimated articulatory features obtained through acoustic-to-articulatory inversion. Specifically, we show that augmenting MFCCs with the estimated articulatory features also significantly enhances the speaker verification performance. Our future works cover investigating better inversion methods that can maximize the inter-speaker articulatory variations as well as applying the proposed MFCC-estimated-articulation feature to the NIST SRE data sets with the state-of-the-art methods.

6. References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[2] P. Kenny, G. Boulianne, P. Dumouchel, and P. Ouellet, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] M. Li, A. Tsiartas, M. V. Segbroeck, and S. Narayanan, "Speaker verification using simplified and supervised i-vector modeling," in *Proceedings of ICASSP*, 2013.

[6] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proceedings of INTERSPEECH*, vol. 4, 2006, pp. 1471–1474.

[7] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of ICCV*, 2007, pp. 1–8.

[8] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *Proceedings of ICASSP*, 2011, pp. 4828–4831.

[9] M. Li, C. Lu, A. Wang, and S. Narayanan, "Speaker verification using lasso based sparse total variability supervector and probabilistic linear discriminant analysis," in *NIST Speaker Recognition Evaluation workshop*, 2011.

[10] M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in *Proceedings of INTERSPEECH*, 2011, pp. 4548–4551.

[11] M. Li and S. Narayanan, "Robust talking face video verification using joint factor analysis and sparse representation on GMM mean shifted supervectors," in *Proceedings of ICASSP*, 2011.

[12] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[13] G. E. Peterson and H. L. Barney, "Control methods used in a study of vowels," *Journal of the Acoustical Society of America*, vol. 24, pp. 175–184, 1952.

[14] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton & Co., 1960.

[15] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[16] K. Stevens, *Acoustic Phonetics*. Cambridge, MA.: MIT Press, 1998.

[17] A. Lammert, M. Proctor, and S. Narayanan, "Morphological variation in the adult hard palate and posterior pharyngeal wall," *Journal of Speech, Language and Hearing Research*, in press.

[18] A. Lammert, M. Procto, and S. Narayanan, "Interspeaker variability in hard palate morphology and vowel production," *Journal of Speech, Language and Hearing Research*, in revision.

[19] A. Lammert, M. Proctor, A. Katsamanis, and S. Narayanan, "Morphological variation in the adult vocal tract: A modeling study of its potential acoustic impact," in *Proceedings of INTERSPEECH*, 2011.

[20] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proceedings of ICASSP*, vol. 1, 1996, pp. 346–348.

[21] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proceedings of ICASSP*, vol. 1, 1996, pp. 353–356.

[22] J. Perkell, "Articulatory processes," in *The Handbook of Phonetic Sciences*, W. Hardcastle and J. Laver, Eds. Oxford: Blackwell, 1997, pp. 333–370.

[23] C. Mooshammer, P. Perrier, C. Geng, and D. Pape, "An EMMA and EPG study on token-to-token variability," *AIPUK*, vol. 36, pp. 47–63, 2004.

[24] J. Brunner, S. Fuchs, and P. Perrier, "The influence of the palate shape on articulatory token-to-token variability," *ZAS Papers in Linguistics*, vol. 42, pp. 43–67, 2005.

[25] J. Brunner, S. Fuchs, P. Perrier, *et al.*, "On the relationship between palate shape and articulatory behavior," *Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3936–3949, 2009.

[26] S. Dart, "Articulatory and acoustic properties of apical and laminal articulations," in *UCLA Working Papers in Phonetics*, I. Maddieson, Ed., 1991, no. 79.

[27] M. Honda, A. Fujino, and T. Kaburagi, "Compensatory responses of articulators to unexpected perturbation of the palate shape," *Journal of Phonetics*, vol. 30, pp. 281–302, 2002.

[28] M. Thibeault, L. Ménard, S. Baum, G. Richard, and D. McFarland, "Articulatory and acoustic adaptation to palatal perturbation," *Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2112–2120, 2011.

[29] M. Li, J. Kim, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, "Automatic classification of palatal and pharyngeal wall shape categories from speech acoustics and inverted articulatory signals," in *Submitted to INTERSPEECH*, 2013.

[30] D. Pechyony and V. Vapnik, "On the theory of learning with privileged information," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[31] O. Vinyals, Y. Jia, L. Deng, and T. Darrell, "Learning with recursive perceptual representations," in *Advances in Neural Information Processing Systems*, 2012, pp. 2834–2842.

[32] A. Toutios and K. Margaritis, "A rough guide to the acoustic-to-articulatory inversion of speech," in *6th Hellenic European Conference of Computer Mathematics and its Applications, HERCMA-2003*, 2003.

[33] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121, p. 723, 2007.

[34] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent, "X-ray microbeam speech production database," *Journal of the Acoustical Society of America*, vol. 88, no. S1, pp. S56–S56, 1990.

[35] P. K. Ghosh and S. S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[36] P. Ghosh and S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *Proceedings of ICASSP*, 2011.

[37] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[38] J. Kim, A. Lammert, P. K. Ghosh, and S. S. Narayanan, "Spatial and temporal alignment of multimodal human speech production data: realtime imaging, flesh point tracking and audio," in *Proceedings of ICASSP*, 2013.

[39] A. Wrench, "MOCHA-TIMIT," Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, speech database, 1999.

[40] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-icsi-ogi features for asr," in *Proceedings of ICSLP*, vol. 1, 2002, pp. 4–7.

[41] "The NIST Year 2010 Speaker Recognition Evaluation Plan," http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.