Automatic Classification of Palatal and Pharyngeal Wall Shape Categories from Speech Acoustics and Inverted Articulatory Signals

Ming Li[1]*, Adam Lammert*[1]*, Jangwon Kim*[1]*, Prasanta Kumar Ghosh*[2] *and Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, USA
[2]Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India

mingli@usc.edu, lammert@usc.edu, jangwon@usc.edu, prasantg@ee.iisc.ernet.in, shri@sipi.usc.edu

Abstract

Inter-speaker variability is pervasive in speech, and the ability to predict sources of inter-speaker variability from acoustics can afford scientific and technological advantages. An important source of this variability is vocal tract morphology. This work proposes a statistical model-based approach to classifying the shape of the hard palate and the pharyngeal wall from speech audio. We used principal component analysis for the parameterization of the morphological shape. Analysis using K-means clustering showed that both the palate and the pharyngeal wall shape data group into two major categories. These in turn are used as targets for automatic classification using acoustic features derived at the utterance level with OpenSmile and at the model level using GMM based posterior probability supervectors. Since articulatory motions are dependent on morphological shape, the model uses estimated articulatory features on top of speech acoustics for improving the classification performance. Experimental results showed 70% and 63% unweighted accuracy for binary classifications of palate and pharyngeal wall shapes in the rtMRI database, respectively, and 63% for the palate shape on the X-Ray Microbeam database.

Index Terms: speech production, vocal tract morphology, acoustic-to-articulatory inversion, speaker recognition

1. Introduction

Issues in speech research often center on inter-speaker acoustic variability. For example, this variability presents challenges for combining speech data from various speakers (requiring speaker normalization [1, 2, 3]). But this inherent variability also provides opportunities to differentiate speakers based on their speech [4, 5]. Regardless of the specific motivation, whether scientific or technological, the ability to understand sources of inter-speaker variability and to predict those sources from acoustics can afford a variety of advantages.

An important source of inter-speaker variability in speech acoustics is the morphology (physical structure) of the vocal apparatus. It has long been understood, for instance, that vocal tract length is a key source of variability in vowel acoustics, with longer vocal tracts resulting in lower formant frequencies [6, 7, 8, 9]. In order to compensate for these differences, automatic speech recognition (ASR) commonly makes use of vocal tract length normalization (VTLN), which has been shown to provide significant gains in system performance [1, 2, 3]. Successful estimation of other morphological patterns may also be useful for handling this inter-speaker variability.

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

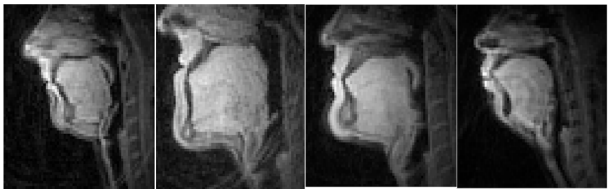
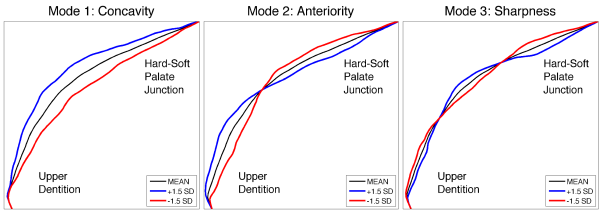


Figure 1: *Vocal tract rtMRI images from different subjects*

The current study focuses on the morphology of the hard palate and posterior pharyngeal wall. Fig.1 shows magnetic resonance images of the vocal apparatus of four different subjects illustrating this variability. These two structures determine much of the vocal tract's morphology on account of their large size and relatively small movements. Although the variety and extent of morphological variation in these structures is well-understood [10], their role in acoustic variability is less understood. Several studies have attempted to reveal the role of palatal morphology in articulatory variability. Speakers with flat palates exhibit less articulatory variability during vowel production than speakers with highly domed palates [11, 12, 13, 14]. Articulation of coronal fricatives is also influenced by palate shape, including apical vs. laminal articulation of sibilants [15], as well as jaw height and the positioning of the tongue body [16, 17].

Differences in hard palate and pharyngeal wall morphology have the potential to alter the resonant properties of the vocal tract and thereby cause acoustic variability [18]. Therefore, this paper focuses on the possibility of automatically characterizing hard palate and pharyngeal wall morphology from speech acoustics. To our best knowledge, there has not been a study on this topic. However, indications are that those difference may not be abundantly evident in the acoustics because speakers adjust their lingual articulation in compensation [19, 20], making estimation of these characteristics from acoustics a very difficult task. Thus, inverted articulatory features are considered, as well, especially in light of the strong evidence showing the previously-mentioned influence of morphology on articulation.

Since we can not acquire real articulation data for general ASR or speaker recognition applications, we applied an exemplar-based acoustic-to-articulatory inversion method [21] to generate the estimated articulatory signal for this work. Inter-speaker variations could be projected into the intra-speaker variabilities of the exemplar speaker when he/she is asked to mimic different speakers' pronunciations. Furthermore, the same MFCC features could be employed for both ASR and



(a) 51% of var. (b) 25% of var. (c) 10% of var.

Figure 2: *Statistical analysis of hard palate shape variations in rtMRI data. Black lines show the mean palate shape, with blue/red lines showing the shape at +/- 1.5 st. dev. from the mean. Concavity differences (a) have been studied, but other important modes include anteriority of the palatal dome (b) and sharpness of the dome(c). (Figure reproduced from [20])*

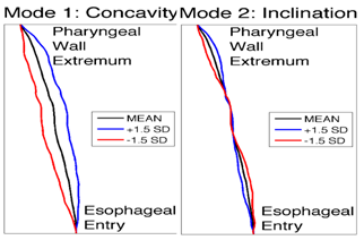


Figure 3: *Statistical analysis of rear pharyngeal wall shape variations in rtMRI data. (Figure reproduced from [20])*

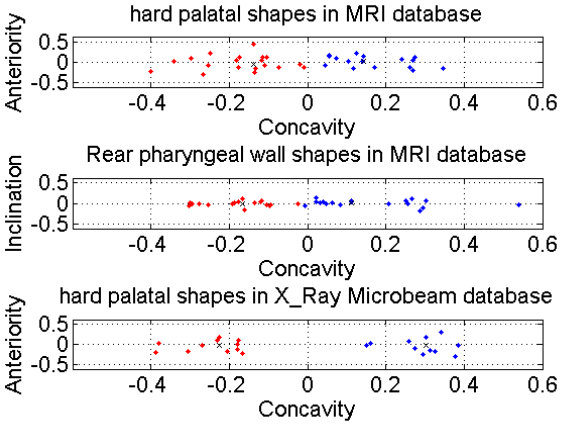


Figure 4: *K-mean clustering of morphology shapes cross all the speakers in each database.*

speaker recognition, therefore the inter-speaker variations may also leak into the inverted articulatory signals through MFCCs. In [22], we showed that the exemplar-based articulatory inversion results, especially mean and variances, still carry inter-speaker variations.

Although the inverted articulatory features are also generated from speech signals, we can show that adding this new information (articulation-acoustics mapping learned from the exemplar's data) on top of MFCCs can still enhance the performance. Theoretical supports from machine learning fields are provided in [23, 24]. Practically, this concatenation based speech-articulatory feature level fusion has been reported to increase the performance of ASR [25, 26] and speaker recognition [22] significantly. In this work, we show that by utilizing information from both speech and inverted articulation, the morphology recognition results is also enhanced.

2. Data

As our evaluation databases, we utilized real-time Magnetic Resonance Imaging (rtMRI) database [27] with synchronized, denoised audio [28] collected in-house, as well as data from the X-Ray Microbeam (XRMB) Speech Production Database [29]. These databases both provide measures of vocal tract morphology in addition to speech recordings from multiple speakers.

2.1. Real-Time MRI Data

The rtMRI database utilized in this study consisted of data from 36 healthy adult with no reported history of speech, language, or hearing pathology. Ages of subjects in this database range between 19 and 37 (mean 27.0, st. dev. 4.3 years). The database also comprised individuals from a variety of language backgrounds: 22 American English speakers, 8 German speakers, 5 Mandarin speakers, and 1 Hindi speaker. All subjects were recorded speaking their native language. This database has previously been used for analysis of morphological variation in the speech production apparatus [20]. Utterances consisted of mostly continuous speech, with an assortment of read passages and spontaneous speech, along with a small number of isolated tokens. The combined total duration of all recorded speech was 58 minutes 4.2 seconds (mean 1 minute 37 seconds per subject).

Data were acquired at Los Angeles County Hospital on a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha, WI) and a custom 4-channel upper airway receiver coil array was used. A 13-interleaf spiral gradient echo pulse sequence was used. Video sequences were reconstructed using a sliding window technique to produce a video rate of 23.18 frames per second with a spatial resolution of 68 pixels $\times$ 68 pixels. Further details can be found in [27, 20].

From Fig.1, we can observe that significant vocal tract morphology variations exist cross different speakers. Based on the statistical analysis of hard palate shape variations in rtMRI data [10], it is shown in Fig.2 that concavity, anteriority and sharpness cover most of the palatal shape variations. These correspond to the height of the palatal dome, the position of the dome's apex in the oral cavity and the angularity of the dome around the apex, respectively. Similarly, we can find concavity and inclination play the most important roles for rear pharyngeal wall shape variations [10] in Fig.3.

2.2. X-Ray Microbeam Data

We also used the X-Ray Microbeam Speech Production Database [29] for evaluation. Compared to the rtMRI database, XRMB data has more speech utterances and no MRI scanner noise involved which enables the acoustic-to-articulatory inversion experiments. For each speaker in the database, the shapes of the hard palate and real pharyngeal wall are approximated by 15 and 2 hand-labeled coordinates, respectively [29]. Due to there are only two discrete measure points for rear pharyngeal wall and the head orientation of subjects were not controlled, we only perform classification for the palatal shape. Furthermore,

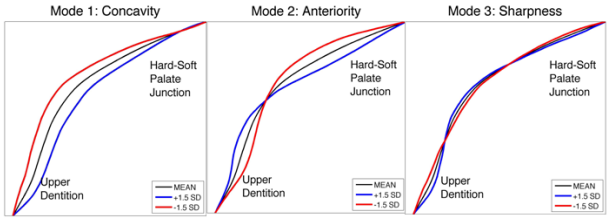


Figure 5: *Statistical analysis of palate shape variations in XRMB data explaining 56%, 35% and 4% of the total variance.*

since there are only limited amount of short duration training data and the system performance is not good enough to accurately regress the concavity value, we perform binary classification on the clustered categorical labels instead. The statistical analysis of palatal shape data and the k-mean clustering result are demonstrated in Fig.5 and Fig.4, respectively, showing this inherent grouping. It is worth noting that the morphology shape analysis illustrated in Fig.2 and 3 was proposed and reported for MRI data in [10, 20]. In this work, we applied the same method on the XRMB database and achieved very similar results as shown in Fig.5, lending further support to the findings in [10]. We selected read speech data (citation words and sentences) from sessions 1 to 101 for each speaker from JW11 to JW63 which amounted to a total of 4034 utterances from 46 speakers with an average duration of 5.72 seconds. Note that we excluded speech sessions that involved speaking in different styles (such as fast or slow speech, emphasized speech, or stimuli that involved diadokinesis). We also omitted speaker sessions that had to be repeated, as well as those which were found to contain severe tracking errors, as detailed in the XRMB Manual [29].

Since the morphology shapes may not be as accurate as the rtMRI data (relative positions of palate markers are not identical), we removed the speakers with concavity values between -0.15 and +0.15 which gave us 1893 utterances from 22 speakers for the binary classification. The advantage of XRMB data is that the speech is relatively clean compared to the rtMRI data which enable us to perform articulatory inversion using models trained on clean speech database. Moreover, the morphology inversion mapping that we studied here could potentially be directly applied to general ASR or speaker recognition applications due to smaller channel mismatch. It is worth noting that the vocal tract morphology variations are not correlated with the gender information [20, 29].

3. Method

We clustered the continuous PCA coefficients of morphology shapes into two broad shape classes and performed a binary classification task. The reason to skip regression is because it requires higher accuracy and more data for training which is not satisfied in this study. From the K-mean clustering results demonstrated in Fig.4, we can see that this binary classification is actually mostly for concavity discrimination which might be because concavity covers more than 50% of the variance. In this work, we just focus on the concavity patterns since it covers the largest covariance in Fig.2 and Fig.5. We performed leave one speaker out cross validation (testing on one speaker's data and training on all other speakers' data with rotation for each speaker) on the palate and pharyngeal wall concavity binary classification using support vector machine (SVM). The two kinds of speech data input vectors are Open Smile supervector (Sec3.1) and UWPP supervector (Sec3.2). In addition to direct speech acoustic features, we also consider articulatory features, obtained by inversion (Sec 3.3), for classification. LibSVM toolkit [30] with RBF kernel was adopted for classification.

3.1. Utterance level features: Open Smile supervector

For generating the speech features for the morphology characterization problem, we resort to using global utterance-level features. Specifically, we use the Open Smile toolkit [31, 32] with the config parameter provided by the 2010 Paralinguistic Challenge for generating the utterance level feature vector which covers various speech features and their functionals, such as MFCC, line spectral pairs frequency, voicing probability, F0, F0 envelop, jitter, and shimmer, etc.

3.2. Model-based features: UWPP supervector

Inspired by recent advances in speaker age/gender modeling, we also propose the use of UBM weight posterior probability (UWPP) supervector to capture the distinct short term speech spectral characteristics of different speakers that can be attributable to the shape variability of interest. After voice activity detection (VAD), non-speech frames were eliminated and cepstral features were extracted. A 25ms Hamming window with 10ms shifts was adopted. 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 normalization were performed to normalize the MFCC features to zero mean and unit variance on a per utterance basis. Wiener filtering [33] was adopted for XRMB data before VAD to reduce stationary artifact noises.

For each utterance in the datasets, UWPP feature extraction is performed on the Universal Background Model (UBM), trained on a population of speakers. Given a frame-based MFCC feature x_t and the GMM-UBM λ with M Gaussian components ($\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, \cdots, M$), the posterior probability is calculated as follows:

$$P(\lambda_i|x_t) = \frac{w_i p_i(x_t|\mu_i, \Sigma_i)}{\Sigma_{j=1}^{M} w_j p_j(x_t|\mu_j, \Sigma_j)}. \quad (1)$$

This posterior probability can also be considered as the fraction of this feature x_t coming from the i^{th} Gaussian component which is also denoted as partial counts. The UWPP supervector is defined as follows:

$$UWPP = [b_1, b_2, \cdots, b_M], b_i = \frac{1}{T}\Sigma_{t=1}^{T} P(\lambda_i|x_t) \quad (2)$$

The reason to adopt UWPP supervector in this study is also due to its relatively small dimension (M) and good generalization capabilities for short duration utterances [32, 34].

3.3. Subject-independent inversion

This section provides a brief description of a subject-independent acoustic-to-articulatory inversion method used in this study. We used generalized smoothness criterion for acoustic-to-articulatory inversion [35] under speaker-indepedent inversion setting [21]. In this setting, a probability feature vector (PFV) is considered instead of the MFCC feature vector for representing the acoustic information; PFV is computed from each MFCC feature vector by computing the

Table 1: Classification result for rtMRI data

methods	Hard palate			
decision level	Utterance		Speaker	
Accuracy type	UA	WA	UA	WA
Open Smile	61	61	69	69
UWPP MFCC	60	61	64	65
Open Smile and UWPP Fusion			70	71
methods	Pharyngeal wall			
decision level	Utterance		Speaker	
Accuracy type	UA	WA	UA	WA
Open Smile	58	59	59	60
UWPP MFCC	58	59	61	62
Open Smile and UWPP Fusion			63	65

probability corresponding to each of the 40 clusters of a general acoustic model as was done in [21]. The generalized smoothness criterion represents variabilities in acoustic space and is built using TIMIT training data [36]. In the subject-independent inversion, the acoustic-articulatory training data is used from only one exemplar speaker and the acoustics of a given (arbitrary) test subject is matched to that of the exemplar using the Euclidean distance measure of the corresponding PFVs. generalized smoothness criterion based optimization ensures the estimated articulatory trajectories to be smooth to a required degree while ensuring PFV-based acoustic similarity between the exemplar and the test speakers. Thus the cost function in generalized smoothness criterion is a weighted sum of two terms representing these two criteria. In our experiments, the acoustic-articulatory data of the exemplar speaker is divided into 5-fold. The hyperparameters of the optimization including the weighing factor, degree of smoothness are optimized on one fold and the rest 4-folds are used for training in inversion.

We use tract variables for representing articulatory space of the exemplar similar to the ones computed in [21]. A female exemplar is chosen from the Electromagnetic articulography (EMA) database collected at the University of Southern California [37]. This database includes speech audio (at 22050-Hz sampling rate) and its parallel recording of six flesh-point sensor positions (at 100-Hz sampling rate) of 460 MOCHA-TIMIT [38] English sentences (about 69 minutes) read by a native speaker of American English. We computed nine vocal tract variables, including lip aperture, lip protrusion, jaw opening, the constriction degrees and constriction locations of tongue tip, tongue blade and tongue dorsum. Constriction location of the each tongue sensor is computed first by projecting the tongue sensor position on the corresponding palate region represented by a straight line, which is chosen by visual inspection of the tongue sensors movements of the exemplar's entire tongue sensor data. After projection, the distance from a fixed point on the straight line is used as the tongue sensor constriction location. More details about the inter-speaker variations of the inverted articulatory signals are provided in [22].

4. Results

Table 1 presents the binary classification results for both hard palate and pharyngeal wall concavity from speech using rtMRI data (Sec2.1). The labels 'Utterance' and 'Speaker' in Table 1 and 2 denote the results of utterance- or speaker-level classification (i.e., one speaker's utterances share a single decision by majority voting). Classification accuracies are reported both as traditional accuracy (weighted accuracy, WA) and unweighted average recall (unweighted accuracy, UA) of these two classes to better compensate for imbalance between classes [32]. The results on XRMB database are shown in Table 2.

Table 2: Classification result for palate in XRMB data

methods	Hard palate	
decision level	Speaker	
Accuracy type	UA	WA
UWPP MFCC	60	66
UWPP MFCC+Inverted Articulation	63	69

5. Discussion

Experiments on rtMRI data show that the Open Smile and UWPP systems were able to classify palatal concavity at 19% and 14% above the 50% chance baseline (Sign and binomial test one-tail $p < 0.0001$), respectively, on a per-speaker basis and these systems were also able to classify pharyngeal wall concavity at 9% and 11% above the baseline (Sign and binomial test one-tail $p < 0.0039$). Score level fusion of these two systems can further enhance the performance.

Notable differences occur in the ability to accurately estimate hard palate shape and pharyngeal wall shape (10%). In general, it is expected that different aspects of morphology may be differentially difficult to estimate, either because they have different potential to impact the acoustics or because speakers compensate for certain morphological variations by adapting their articulation in pursuit of some acoustic speech target. However, there is evidence to suggest that morphological differences in the posterior pharyngeal wall have just as much potential to impact vowel acoustics as hard palate shape [18]. This difference is even more puzzling because articulatory compensation in the pharynx is likely more difficult than in the oral cavity because, without the dexterous tongue tip to employ, there are effectively fewer degrees of freedom in the pharynx. This reasoning may only apply to vowels, however. Morphological differences probably also have the potential to impact the acoustics of stops and fricatives which, in the languages considered here, occur near the palate in overwhelming proportions.

The results on the XRMB database are similar to those on rtMRI data, especially in terms of WA, but with a slightly lower UA. This lower accuracy is very surprising, given the relatively clean audio associated with the XRMB data – as opposed to the denoised rtMRI audio. However, in XRMB data, the morphology shapes may not be as accurate as the rtMRI data (relative positions of palate markers are not identical) which could result in the lower accuracy. The idea that inverted articulatory information has the potential to improve the performance of morphological inversion is supported by the XRMB experiments. By concatenating the inverted articulatory features with MFCCs together, classification accuracies were increased by 3% which matches with the results in [22]. In the future work, we would train the inversion model with multiple subjects' data as well as study more accurate inversion methods.

6. Conclusion

Future works include the continued collection of large scale morphological databases using rtMRI with denoised audio, but also with clean speech data from the same speakers. Phoneme-

specific modeling will also be important for establishing which vocal tract postures provide the best information about morphological characteristics. We also plan to apply the estimated morphology information for supervised I-vector based speaker recognition [5]. We also plan to study vocal tract normalization methods to improve ASR using more detailed morphology information than just vocal tract length.

7. References

[1] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1996, pp. 346–348.

[2] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1996, pp. 353–356.

[3] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1996, pp. 339–341.

[4] 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.

[5] M. Li, A. Tsiartas, M. V. Segbroeck, and S. Narayanan, "Speaker verification using simplified and supervised i-vector modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[6] 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.

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

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] J. Brunner, P. Hoole, and P. Perrier, "Articulatory optimisation in perturbed vowel articulation," in *Proceedings of the International Congress of Phonetic Sciences*, 2007.

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

[21] P. Ghosh and S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011.

[22] M. Li, J. Kim, P. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on speech and inverted articulatory signals," in *submitted to INTERSPEECH*, 2013.

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

[24] 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.

[25] 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.

[26] 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.

[27] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the Acoustical Society of America*, vol. 115, p. 1771, 2004.

[28] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans (l)," *Journal of the Acoustical Society of America*, vol. 120, no. 4, p. 1791, 2006.

[29] 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.

[30] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[31] F. Eyben, M. Wollmer, and B. Schuller, "Openear introducing the munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.

[32] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, 2012.

[33] 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 *Proc. ICSLP*, vol. 1, 2002, pp. 4–7.

[34] M. Li, C.-S. Jung, and K. J. Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition," in *Proc. Interspeech*, 2010.

[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] 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.

[37] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

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