Proc APSIPA 2012
Hollywood, CA

Exploiting speech production information for automatic speech and speaker modeling and recognition – possibilities and new opportunities

Vikram Ramanarayanan*, Prasanta Kumar Ghosh†, Adam Lammert* and Shrikanth S. Narayanan*
* Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA
† IBM India Research Lab, New Delhi, India

***Abstract*—We consider the potential for incorporating direct, or inferred, speech production knowledge in speech technology development. We first review the technologies that can be used to capture speech articulation information. We discuss how meaningful (speech and speaker) representations can be derived from articulatory data thus captured and further how they can be estimated from the acoustics in the absence of these direct measurements. We present some applications that have used speech production information to further the state of the art in automatic speech and speaker recognition. We also offer an outlook on how such knowledge and applications can in turn inform scientific understanding of the human speech communication process.**

I. Introduction

Any speech or speaker modeling procedure, in order to be effective, must choose a structure for the model that reflects the structure of the underlying physical system [1]. For example, automatic speech recognition (ASR) can benefit from knowledge of the coordination of the vocal tract articulators and the resulting acoustics; this can help reduce apparent token-to-token variability so that general pattern recognition algorithms have less work to do [2]. In addition, speakers exhibit substantial differences in many aspects of their individual vocal tract morphology, all of which have the potential to alter acoustic output or force speakers to adjust their articulation in compensation. Incorporating such knowledge into speaker modeling could likewise improve speaker recognition performance.

A perennial challenge however has been access to realistic and useful human speech data. Table I lists current state-of-the-art articulatory measurement techniques and their relative advantages and disadvantages. Techniques that have been used to measure articulation include x-ray microbeam (XRMB) [4], electropalatography (EPG), electromagnetic articulography (EMA) [5] and ultrasound [3]. These techniques, although some are invasive, are able to capture articulatory information at high sampling rates. However, none of these modalities offer a complete a view of all vocal tract articulators at a sufficiently high spatial resolution, which is important for studying vocal tract posture. More recently, developments in real-time MRI have allowed for an examination of shaping along the entirety of the vocal tract during speech production and provide a means for quantifying the choreography of the articulators, including structural/morphological characteristics of speakers in conjunction with their articulation and acoustics [7]. However, current rt-MRI protocols have an intrinsically lower frame rate than the other modalities.

II. Representations

Enhanced articulatory representations derived from articulatory data have the potential to inform work in both phonetic and phonological theory, as well as speech and speaker modeling. The problem of seeking a set of representations of the human speech production process can be approached in either a knowledge-driven or a data-driven manner. An example of the former from the linguistics (and more specifically, phonology) literature is the framework of Articulatory Phonology [8] which theorizes that the act of speaking is decomposable into units of vocal tract actions termed "gestures." Under this gestural hypothesis, the primitive units out of which lexical items are assembled are constriction actions of the vocal organs. So in this framework, a simple set of linguistically-meaningful primitive representations are so-called 'constriction task variables' (or a set of constriction degrees and locations); this is one possible basis set that can be used to characterize the gestural lexicon of a language used in speech planning. One can also derive articulatory representations in a data-driven manner. One can broadly classify these into two types – those obtained directly via articulatory measurements and those that can be estimated from non-articulatory sources (such as the speech signal).

A. Representations obtained through direct articulatory measurements

One straightforward articulatory feature set that may be used comprises the raw articulatory measurements themselves. Take for instance, the raw pellet trajectories obtained using techniques like XRMB or EMA (Table I). Raw pixel intensities from rtMRI and their variation over time can also be used as a feature [9]. Further, coordinative relationships between articulators can be quantified by calculating pixel correlation [10]. However, we may desire to derive more intuitively meaningful representations depending on the application. For instance, these could be outlines or contours that delineate the tongue and vocal tract structures. In the case of XRMB or EMA, contours may be obtained by fitting a smooth spline through all pellet points. However, in the case of MRI and ultrasound,

TABLE I
Articulatory measurement techniques.

Characteristic	XRMB	EMA	Ultrasound	EPG	rtMRI
Order of typical sampling rate (Hz)	100	500	50 to 300	100	20 to 30
Relative spatial resolution	Low	Low	Medium	High	High
View of vocal tract	Fleshpoints	Fleshpoints	Tongue	Tongue–palate contact	Full view
Supine position?	No	No	No	No	Yes
Invasive?	Yes	Yes	No	Yes	No
Example database (with citation)	Wisconsin x-ray microbeam database [4]	Edinburgh MOCHA database [5]	Haskins HOCUS [3]	Edinburgh MOCHA database [5]	USC MRI–TIMIT database [6]

more involved image processing is required to segment air-tissue boundaries. For example, we have developed a robust tool for unsupervised region segmentation of the upper airway, jaw and supraglottal articulators, which is suited for processing long sequences of MR images. The segmentation algorithm uses an anatomically informed object model, and returns a set of tissue boundaries for each frame of interest, allowing for quantification of articulator movement and vocal tract aperture in the midsagittal plane (see Figure 1). Further details of the region segmentation algorithm may be found in [11]. Once vocal tract contours are obtained, we can further compute other meaningful representations – such as area functions, cross-distances, and other postural variables ([12], [13]). Area functions are obtained by first imposing a semi-polar grid on the midsagittal image of the vocal tract and then finding the intersections between each gridline and the vocal tract contour outlines found earlier. Finally, the distances between the intersection coordinates on each gridline are computed, starting from the lips to the glottis, and use this ordered set of cross-distances as a feature vector to capture vocal tract posture. Note that although elegant, this procedure suffers from one major disadvantage – it is only semi-automatic: one has to manually choose the initial parameters of the semi-polar grid to be fitted to the vocal tract (such as the number of gridlines, spacing between gridlines, gridline orientation angle, to name a few). This also means that there is minimal guarantee that one will be able to compare gridlines at the same position across different subjects. Recently, Ramanarayanan *et al.* [14] proposed a method to automatically derive cross-distances that are computed at points where constrictions are made in the vocal tract during normal speech production, such as the alveolar ridge for coronal stop consonants, or the lips for labial stops. Hence they are conducive to meaningful comparison across subjects. In addition, other meaningful postural features such as the angle of the jaw or the centroid of the tongue can be computed from segmented MRI data.

B. Representations obtained through estimated articulatory measurements

While speech production data acquisition using EMA or MRI technology has opened up possibilities for new research, it is important to note at the same time that the acquisition of such speech production data can be expensive, intrusive or impractical. Recording such data could also be prohibitive and/or uncomfortable to the subjects in many applications. Such scenarios could be good venue for research on acoustic-to-articulatory inversion of inferring veritable articulatory information from representative datasets or associated measurements. In acoustic-to-articulatory inversion (AAI), articulatory (speech production) representations are estimated from the acoustic speech signal. Hence AAI can be useful in cases when it is not possible to directly measure the speech production data.

There are several approaches available for AAI. Essentially, these approaches can be classified into four broad classes - (1) approaches that make use of quantitative models or mathematical relations between acoustic and articulatory space representations; (2) analysis-by-synthesis or codebook look-up approaches [15]; (3) Artificial Neural Network-based approaches [16]; and (4) statistical approaches ([17], [18], [19], [20]). A classic example of the first approach is where the vocal tract shapes are estimated from the speech acoustics using mathematical relation between the formant frequencies in vowel sounds and the area function of the vocal tract [21]. Other examples of approaches in this category use Maeda's articulatory model [22], or linear-prediction based approach [23]. In the codebook look-up approach, the articulatory space is quantized and the corresponding acoustic features are synthesized to form a codebook of acoustic/articulatory vector pairs. For example, in [15] the codebook is represented in the form of a hierarchy of hypercubes and each hypercube represents a region of the articulatory space in which the mapping is linear. In the neural network based approach, the parameters of the networks are trained to get a nonlinear mapping between articulatory and acoustic parameters. These approaches are most useful when the articulatory space is represented by means of abstract linguistics-derived parameters. Unlike the previous approaches, the final class of methods, based on stochastic modeling and statistical inference methods, is appropriate when there is access to parallel corpora of acoustics and articulation from which mappings can be learned. There have been several approaches to approximate this mapping function, for example using a mixture density network [17] or using dynamical system modeling (Kalman filtering) [18]. Other methodological improvements have been proposed to incorporate increasing realism and robustness to the techniques. For example Ghosh et al. proposed a generalized smoothness criterion to ensure that the estimated articulatory trajectories are smooth to the required degree [19].

In general, statistical approaches that have been proposed have been subject-dependent inversion schemes since they

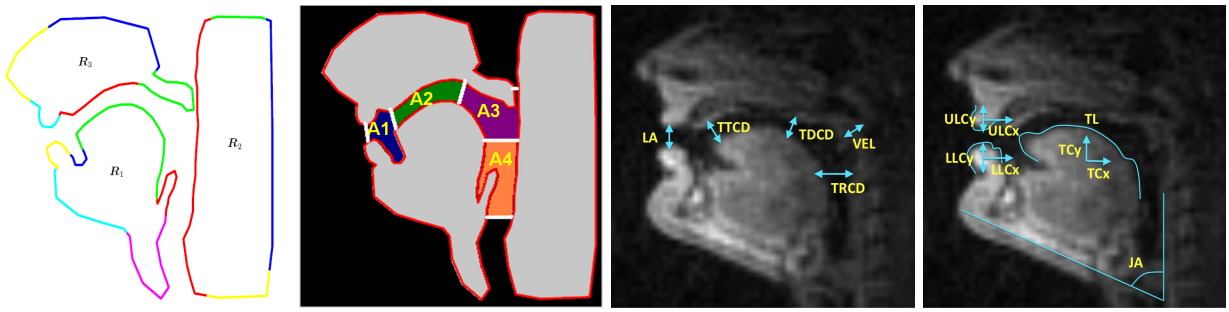

Fig. 1. Examples of meaningful features that can be computed from rt-MRI data. (a) Contour outlines [11]. (b) Meaningful cross-distances can be computed that segment the vocal tract into areas A1-A4 [14]. (c) Cross-distances in more detail (lip aperture (LA), velic aperture (VEL), and constrictions of the tongue tip (TTCD), tongue dorsum (TDCD) and tongue root (TRCD). (d) Articulatory posture variables – jaw angle (JA), tongue centroid (TC) and length (TL), and upper and lower lip centroids (ULC and LLC).

perform inversion well only on the subject's data it is trained with; a basic reason for this is the inherent articulatory-acoustic variability across talkers in their production details. To overcome this subject dependence, Ghosh et al. have proposed a subject-independent inversion scheme [24] where the acoustic variability during testing is normalized using a general acoustic space created from a pool of general talker speech acoustic data. Hence, when speech data from an arbitrary talker (not seen in training) need to be inverted, the normalization allows for inversion in a "subject-independent" way. Such a subject-independent scheme can perform inversion on any subject's acoustics even though that subject does not belong to the training data.

Thus various approaches for AAI can be used to derive speech production representations from the speech signal, especially in applications where only speech is recorded and no direct speech articulation data are available.

III. Applications

A. Automatic speech recognition (ASR)

There have been several production-oriented approaches to automatic speech recognition using both direct and estimated data as well as using recognition models developed based on speech production knowledge. For example, [25] showed improvement in speech recognition accuracy by combining acoustic and articulatory features from a talker. However, it is not practical to assume the availability of direct articulatory measurements from a talker in real-world speech recognition scenarios. To address this challenge, a number of techniques have been proposed ([26], [27], [28]) where, instead of relying on features from direct articulatory measurements, abstracted articulatory knowledge is incorporated in designing models (e.g., Dynamic Bayesian Networks (DBNs), Hidden Markov Models (HMMs)) which can be gainfully used for automatic speech recognition. A summary of such techniques can be found in [29]. [30], [31] proposed an integrated Bayesian framework for ASR that consists of a hard-wired lexical compilation/representation component (which attempts to generalize ideas of feature overlap proposed by phonological theories so that the acoustic space can be modeled with fewer atomic speech units) as well as a stochastic acoustic mapping component. Multi-steam architectures [32] have been also proposed as an alternative approach where linguistically derived articulatory (or more generally, phonetic) features are estimated from the acoustic speech signal, typically using Artificial Neural Networks (ANN), and then used to either replace or augment acoustic observations in an existing HMM based speech recognition system.

In the context of articulatory data-driven approach for recognition, [33] has used estimated articulatory features obtained through subject-independent AAI to address the challenge of unavailability of direct speech production data during speech recognition. Using subject-independent AAI II-B, articulatory features for any talker is estimated using an inversion scheme trained with parallel acoustic-articulatory data from an exemplary subject. Using this technique, the authors were able to show an absolute improvement of around 2% on a HMM-based phone recognition task using the TIMIT corpus. To understand the reason for this improvement, note that the estimated articulatory features could be interpreted as dynamical features derived from the sequence of short-time acoustic feature vectors [19]. Dynamical features could provide information about phonetic classes that is complementary to that provided by short-time features, and could hence lead to a boost in recognition accuracy. New advances in measuring large scale data sets using modalities such as rtMRI, EMA, etc., offer the opportunity to further develop this exemplar-based speech recognition paradigm possibly across multiple languages.

B. Speaker recognition

Understanding the interplay of vocal tract structure, articulation and acoustics has technological applications for automatic speaker recognition. Vocal tract length normalization is one example of morphological knowledge that has already provided performance benefits to automatic speech and speaker recognition (see for example [34]). Possibilities exist for providing normalization of the acoustic signal for other structural differences that impact a variety of phonemes [35]. An essential component of this normalization, in terms of making if practically useful, is to accurately predict morphological characteristics of a speaker from the acoustic signal

(i.e., morphological inversion). Predictions of this kind may subsequently lead to applications in speaker recognition. These features will be unique to an individual speaker, making them ideal for biometric applications.

C. Articulatory modeling

Speech production data can facilitate research in articulation models describing the full vocal tract shape dynamics. For example, [9] investigated the application of statistical graphical models that can capture the spatio-temporal dependencies between various articulators in a data-driven manner. This study indicates that if we combine (a) an explicit multistream transcription with (b) appropriate techniques for extracting articulatory time-functions along with (c) the appropriate statistical models, we are well-positioned to derive phonological information directly from articulatory data. Recently [36] proposed a modeling framework to validate different articulatory representations using articulatory recognition, which affords an understanding of the usefulness of a given representation in analyzing speech articulation.

IV. Scientific impact

A. Articulatory primitives and motor control

Consider the case of speech motor control. One popular theory of motor control is the inverse dynamics model, i.e., in order to generate and control complex behaviors, the brain needs to explicitly solve systems of coupled equations. [37] and [38] instead argue for a less computationally complex viewpoint wherein the central nervous system uses a set of '*primitives*' to "solve" the inverse dynamics problem. Articulatory movement primitives may be defined as a dictionary or template set of articulatory movement patterns in space and time, weighted combinations of the elements of which can be used to represent the complete set of coordinated spatio-temporal movements of vocal tract articulators required for speech production. We recently proposed an algorithm to automatically extract such primitives from speech articulation data [39].

Consider further the case of coarticulation in speech, where the position of an articulator/element may be affected by the previous and following target [40]. Using the idea of motor primitives, we can explore how the choice, ordering and timing of a given movement element within a well-rehearsed sequence can be modified through interaction with its neighboring elements (coarticulation). For instance, through a handwriting-trajectory learning task, [41] demonstrate that extensive training on a sequence of planar hand trajectories passing through several targets results in the coarticulation of movement components, and in the formation of new movement primitives.

B. Link between speech production and perception

There have been several well-known hypotheses regarding the relation between production and perception systems in human speech communication ([42], [43]). Quantitatively modeling these relationships in order to develop better models of ASR and speaker recognition is a challenging task and has not been addressed well by researchers. The availability of rich speech production data has opened horizons for addressing these challenging research questions in a data-driven manner. There are a few advances in this regard. For example, using mutual information as a metric, Ghosh et al. [44] have shown in a data-driven manner that the non-uniform auditory filterbank in the human ear (receiver) is optimum in providing least uncertainty in decoding articulatory movements in the human speech production system (transmitter). This is an exciting finding since it indicates that the design of the filterbank for speech recognition systems needs to be optimally designed with respect to the characteristics of the speech production system. The same authors also proposed an exemplar-specific model for speech recognition using speech production knowledge from an exemplar speaker; this is another attempt to exploit the production-perception relationship. Such a model has also been shown to be well-suited to investigate the effect of language mismatch between the talker and exemplar in a cross-language ASR application [33]. More such computational models need to be developed to understand the effect of speaker dependence, language effect, pathologies and paralinguistic features in speech and speaker recognition tasks, particularly to discover robust recognition models.

C. Speaker morphology and its acoustic impact

Examinations of the interplay between vocal tract structure, articulation and acoustics have already provided insights into essential issues in speech production research, including longstanding questions related to inter-speaker variability and the nature of speech production goals. Structure dictates the space of vocal tract shapes, the ways which those configurations can be achieved and the resonant and other acoustical properties of the system. Morphological variation is therefore a potential source of variability in both the articulatory and acoustic domains, and has the potential to explain the pervasive production variations that are observed across speakers (see for example [45]). Moreover, the extent to which the speech controller minimizes production in articulation versus acoustics within a single speaker can provide evidence in favor of specific definitions of the goals of speech production [46].

V. Outlook

We have presented some potential applications of speech production knowledge in speech technology design. Meaningful representations derived from articulatory data have tremendous potential to further the state-of-the-art in automatic speech and speaker recognition as well as advancing speech science. Further, speech signals inherently carry paralinguistic and extralinguistic information that are interwoven together with their linguistic content – these could be affective, personal or transmittal in nature. While human listeners do an excellent job of processing these different sources of information, speech recognition systems are still lacking in this regard. A major reason for this performance gap can be attributed to robustness issues, i.e., the limited ability of current systems to

successfully tease apart different sources of variability in the speech signal. In this paper we have argued that incorporating knowledge of speech sound production as well as emotional affect, prosody, and speaker morphology can inform modeling of these different sources of variability. This, we believe, will in turn go a long way in improving the state of the art in speech technology.

ACKNOWLEDGMENTS

The authors' work described in this paper was supported by NIH grants DC007124 and DC03172, the USC Imaging Sciences Center, the LAC-USC hospital and the USC Center for High Performance Computing and Communications (HPCC).

REFERENCES

[1] R. Rose, J. Schroeter, and M. Sondhi, "The potential role of speech production models in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 99, p. 1699, 1996.

[2] R. McGowan, "Knowledge from speech production used in speech technology: Articulatory synthesis," *Haskins Laboratories Status Report on Speech Research*, vol. SR-117/118, pp. 25–29, 1994.

[3] D. Whalen, K. Iskarous, M. Tiede, D. Ostry, H. Lehnert-LeHouillier, E. Vatikiotis-Bateson, and D. Hailey, "The haskins optically corrected ultrasound system (hocus)," *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 3, p. 543, 2005.

[4] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America*, vol. 88, p. S56, 1990.

[5] A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," in *Workshop on Phonetics and Phonology in ASR*, 2000.

[6] S. Narayanan, E. Bresch, P. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time mri articulatory corpus for speech research," in *Proc. 12th Conf. Intl. Speech Communication Assoc. (Interspeech 2011)*, Florence, Italy, 2011.

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

[8] C. Browman and L. Goldstein, "Dynamics and articulatory phonology," *In T. van Gelder and B. Port (Eds.), Mind as motion: Explorations in the dynamics of cognition*, pp. 175–193, 1995.

[9] E. Bresch, A. Katsamanis, L. Goldstein, and S. Narayanan, "Statistical multi-stream modeling of real-time mri articulatory speech data," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] A. Lammert, M. Proctor, and S. Narayanan, "Data-driven analysis of realtime vocal tract mri using correlated image regions," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[11] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 3, pp. 323–338, 2009.

[12] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[13] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," *Speech production and speech modelling*, pp. 131–149, 1990.

[14] V. Ramanarayanan, D. Byrd, L. Goldstein, and S. Narayanan, "Investigating Articulatory Setting-Pauses, Ready Position, and Rest-Using Real-Time MRI," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[15] S. Ouni and Y. Laprie, "Improving acoustic-to-articulatory inversion by using hypercube codebooks," in *Sixth International Conference on Spoken Language Processing*, 2000.

[16] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000.

[17] K. Richmond, "Mixture density networks, human articulatory data and acoustic-to-articulatory inversion of continuous speech." 2001.

[18] S. King and A. Wrench, "Dynamical system modelling of articulator movement." International Congress of Phonetic Sciences, 1999.

[19] P. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, p. 2162, 2010.

[20] A. Lammert, L. Goldstein, S. Narayanan, and K. Iskarous, "Statistical methods for estimation of direct and differential kinematics of the vocal tract," *Speech Communication*, 2012.

[21] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 21, no. 5, pp. 417–427, 1973.

[22] Y. Laprie and B. Mathieu, "A variational approach for estimating vocal tract shapes from the speech signal," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 929–932.

[23] S. Krstulovic, "Lpc modeling with speech production constraints," in *Proceedings 5th Speeech Production Seminar*, 2000.

[24] P. Ghosh and S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4624–4627.

[25] J. Frankel and S. King, "Asr-articulatory speech recognition," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[26] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 22, no. 2, pp. 93–111, 1997.

[27] L. Lee, H. Attias, and L. Deng, "Variational inference and learning for segmental switching state space models of hidden speech dynamics," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–872.

[28] J. Ma and L. Deng, "Target-directed mixture dynamic models for spontaneous speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 1, pp. 47–58, 2004.

[29] E. Mcdermott and A. Nakamura, "Production-oriented models for speech recognition," *IEICE transactions on information and systems*, vol. 89, no. 3, pp. 1006–1014, 2006.

[30] L. Deng and D. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *The Journal of the Acoustical Society of America*, vol. 95, p. 2702, 1994.

[31] L. Deng, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Communication*, vol. 24, no. 4, pp. 299–323, 1998.

[32] F. Metze and A. Waibel, "A flexible stream architecture for asr using articulatory features," in *Seventh International Conference on Spoken Language Processing*, 2002.

[33] P. Ghosh and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, 2011.

[34] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 1, pp. 49–60, 1998.

[35] 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*, 2012.

[36] A. Katsamanis, E. Bresch, V. Ramanarayanan, and S. Narayanan, "Validating rt-MRI based articulatory representations via articulatory recognition," Florence, Italy, Aug 2011.

[37] F. Mussa-Ivaldi, N. Gantchev, and G. Gantchev, "Motor primitives, force-fields and the equilibrium point theory," *From Basic Motor Control to Functional Recovery. Academic Publishing House" Prof. M. Drinov", Sofia, Bulgaria*, pp. 392–398, 1999.

[38] C. Hart and S. Giszter, "A neural basis for motor primitives in the spinal cord," *The Journal of Neuroscience*, vol. 30, no. 4, pp. 1322–1336, 2010.

[39] V. Ramanarayanan, A. Katsamanis, and S. Narayanan, "Automatic data-driven learning of articulatory primitives from real-time mri data using convolutive nmf with sparseness constraints," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[40] D. Ostry, P. Gribble, and V. Gracco, "Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned?" *The Journal of Neuroscience*, vol. 16, no. 4, pp. 1570–1579, 1996.

[41] R. Sosnik, B. Hauptmann, A. Karni, and T. Flash, "When practice leads to co-articulation: the evolution of geometrically defined movement primitives," *Experimental Brain Research*, vol. 156, no. 4, pp. 422–438, 2004.

[42] B. Lindblom, "Role of articulation in speech perception: Clues from production," *The Journal of the Acoustical Society of America*, vol. 99, p. 1683, 1996.

[43] S. Wilson *et al.*, "Listening to speech activates motor areas involved in speech production," *Nature neuroscience*, vol. 7, no. 7, pp. 701–702, 2004.

[44] P. Ghosh, L. Goldstein, and S. Narayanan, "Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures," *The Journal of the Acoustical Society of America*, vol. 129, p. 4014, 2011.

[45] S. Fuchs, R. Winkler, and P. Perrier, "Do speakers' vocal tract geometries shape their articulatory vowel space?" in *8th International Seminar on Speech Production, Strasbourg, France*, 2008.

[46] L. Ménard, J. Schwartz, L. Boë, and J. Aubin, "Articulatory–acoustic relationships during vocal tract growth for french vowels: Analysis of real data and simulations with an articulatory model," *Journal of Phonetics*, vol. 35, no. 1, pp. 1–19, 2007.