INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

VCV Synthesis using Task Dynamics to Animate a Factor-based Articulatory Model

Rachel Alexander, Tanner Sorensen, Asterios Toutios, Shrikanth Narayanan

Signal Analysis & Interpretation Laboratory (SAIL)
University of Southern California, Los Angeles, CA, USA
{rachela, tsorense, toutios}@usc.edu, shri@sipi.usc.edu

Abstract

This paper presents an initial architecture for articulatory synthesis which combines a dynamical system for the control of vocal tract shaping with a novel MATLAB implementation of an articulatory synthesizer. The dynamical system controls a speaker-specific vocal tract model derived by factor analysis of mid-sagittal real-time MRI data and provides input to the articulatory synthesizer, which simulates the propagation of sound waves in the vocal tract. First, parameters of the dynamical system are estimated from real-time MRI data of human speech production. Second, vocal-tract dynamics is simulated for vowel-consonant-vowel utterances using a sequence of two dynamical systems: the first one starts from a vowel vocal-tract configuration and achieves a vocal-tract closure; the second one starts from the closure and achieves the target configuration of the second vowel. Third, vocal-tract dynamics is converted to area function dynamics and is input to the synthesizer to generate the acoustic signal. Synthesized vowel-consonant-vowel examples demonstrate the feasibility of the method.

Index Terms: articulatory synthesis, task dynamics, real-time MRI, factor analysis, articulatory model, coarticulation

1. Introduction

Several attempts have been made in the past to synthesize speech by inferring the dynamics of the area function and simulating the physics of the propagation of sound in the vocal tract [1, 2, 3, 4]. One of the most prominent and comprehensive such attempts is TaDA [5], from Haskins Laboratories, which implements notions from the theories of Articulatory Phonology [6] and Task Dynamics [7] to animate Mermelstein's classic model of the mid-sagittal vocal-tract geometry [8, 9].

Real-time magnetic resonance imaging (rtMRI) has enabled the acquisition of high-speed imaging data from the entire vocal tract in unprecedented volumes [10, 11]. Advances in automatic air-tissue segmentation and statistical factor analysis of the air-tissue boundaries [12] have enabled the development of individualized, speaker-specific, articulatory models. Such factor-based models capture the natural deformations of a given speaker's articulators, in a way that is arguably more realistic compared to generic, speaker independent models such as that proposed by Mermelstein [8].

In previous work [13], we have developed a framework to construct dynamical systems that deform such factor-based models in order to achieve constrictions at various places of articulation. The present work extends that framework to animate such an articulatory model for VCV sequences. The resulting vocal-tract dynamics is then converted to area-function dynamics via a commonplace $\alpha\beta$-model, which drives an articulatory synthesizer to generate speech acoustics.

The articulatory synthesizer we use is a novel MATLAB implementation of the method proposed by Maeda to solve in the time domain a time-varying lumped electrical network, whose elements are functions of the dynamically changing cross-sectional dimensions of the vocal tract, including the glottis [14, 15]. Our synthesizer, accompanied by synthesis results, is available online.[1]

This paper focuses on VCV sequences with oral voiced consonants. Synthesis of unvoiced consonants requires a careful modeling of the temporal coordination between the oral constriction and glottal abduction, which is a goal for future research. Similar considerations apply to the coordination between oral constriction and velopharyngeal opening for nasals.

2. Methods

2.1. Articulatory Model

Mid-sagittal vocal-tract configurations are represented by weighted sums of articulatory parameters obtained from rtMRI data of the mid-sagittal slice [12]. The 460 sentences of the MOCHA-TIMIT set [16] were spoken by the F1 speaker of the USC-TIMIT database [17] and recorded with rtMRI at a rate of 23.18 fps. Images of the mid-sagittal slice from these data underwent an automatic segmentation algorithm [18] to obtain a vocal-tract outline of 184 points on the mid-sagittal xy plane that described 15 anatomical features.

A factor analysis [12] was applied to the coordinates of these points to determine a set of constant factors that correspond to speaker-specific linguistically meaningful articulatory components (Fig. 1). Specifically, there is a factor for the jaw movement, four additional degrees of freedom for the tongue (after removal of the jaw contribution), two additional degrees of freedom for the lips and an independent velum factor. Each vocal-tract configuration is then compactly represented by a vector of weights (or, *control parameters* of the articulatory model) that correspond to the degree of deformation of each factor, and can be used to accurately reconstruct the vocal tract.

2.2. Tract Variables

Constriction degrees were defined in previous work [13] to determine the distance between surfaces of active and passive articulators at the places of articulation specified in Fig. 2. For each vocal-tract configuration, 6 such constriction degrees, or tract variables, were computed as the minimum distance between opposing surfaces within the boundaries of each place of constriction.

Tract variables were related to the articulatory parameters

[1] http://sail.usc.edu/span/artsyn2017 and submitted as supplementary material.

http://dx.doi.org/10.21437/Interspeech.2017-1410

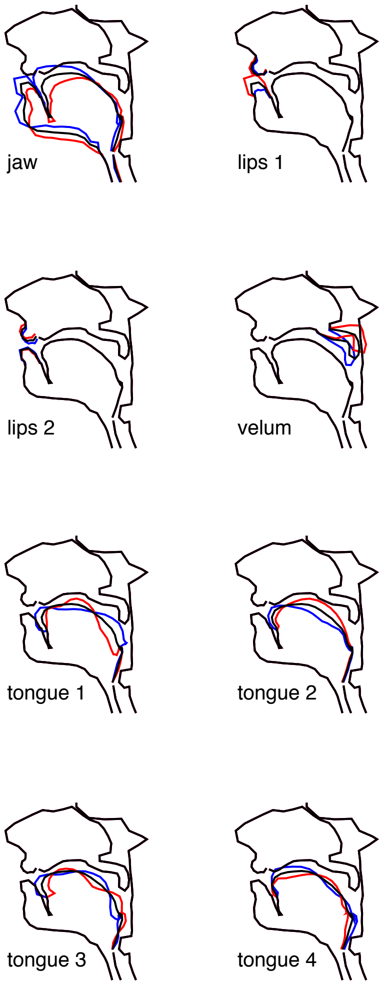

Figure 1: *Factors of the articulatory model. Line segments denote mean and* ±2 *standard deviations*

described above using a locally-linear map defined by a hierarchical clustering procedure, which associates tract variables to the corresponding articulatory parameters at each frame of the rtMRI video, linking constriction degrees to vocal-tract shape. The hierarchical clustering algorithm estimates the forward map [19, 20] $\mathbf{G} : \mathbf{R}^8 \rightarrow \mathbf{R}^6$, which maps articulatory parameters to tract variables. The algorithm computes a tree whose root node is the set of all observed articulatory parameter vectors in the dataset. A k-means subroutine starts at the root and iteratively breaks nodes in two (i.e., $k = 2$). Children in this tree are disjoint subsets of the parent and the union of siblings is the parent. Nodes stop breaking either when a child would contain fewer than 9 articulatory parameter vectors (to prevent rank-deficiency in least squares estimation of $\mathbf{G}$) or when $\mathbf{G}$ maps the articulatory parameters of that node to tract variables in $\mathbf{R}^6$ approximately linearly (i.e., when $\mathbf{G}(\mathbf{w})$ estimates $\mathbf{z}$ with a mean absolute error of less than λ mm, where λ is a free parameter chosen to be 0.24 mm). Within each terminal node, the algorithm uses least squares to estimate $\mathbf{G}$, the Jacobian J of $\mathbf{G}$, and the time derivative $\dot{J}$ of the Jacobian. By change of variables $\mathbf{z} = \mathbf{G}(\mathbf{w})$, $\dot{\mathbf{z}} = J\mathbf{w}$, $\ddot{\mathbf{z}} = J\dot{\mathbf{w}} + \dot{J}\mathbf{w}$, articulatory parameters could be converted to tract variables and vice versa, allowing for inputs and outputs from the dynamical system to be used as vocal-tract dynamics for articulatory synthesis.

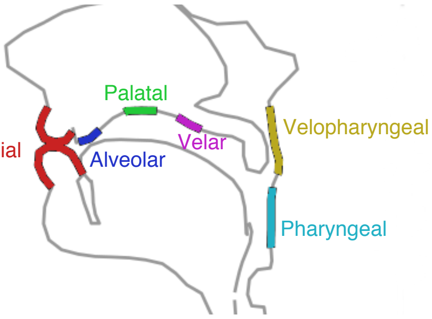

Figure 2: *Locations of constrictions considered in this study. (Bilabial, alveolar, palatal, velar, pharyngeal, and velopharyngeal)*

The forward map $\mathbf{G}$ allows us to describe vowel configurations with articulatory parameters (weight vectors) and to describe the consonant configurations with tract variables (i.e., a target constriction degree at one of the places of articulation in Fig. 2). The linear map allowed these two representations to interface in the dynamical system.

2.3. Dynamical System

The methodology used here was developed on the basis of Task Dynamics [7] to achieve constrictions in the vocal tract by deforming an initial set of parameters to fit a target configuration. A dynamical systems model was produced to achieve a target vocal-tract shape by transforming the articulatory parameters from the data into tract variables using the forward map, and adjusting the degrees of constriction to control vocal-tract deformation [13]. The dynamical system finds an optimal deformation from the initial vocal-tract shape represented by the appropriate articulatory parameters, to the target constriction represented by the appropriate tract variables.

We describe change in the vector $\mathbf{z}$ of tract variables over time with the differential equation $\ddot{\mathbf{z}} = -K(\mathbf{z} - \mathbf{z}_0) - B\dot{\mathbf{z}}$, where $\mathbf{z}_0$ is a vector of 6 tract variable targets and K and B are 6×6 diagonal matrices of stiffness and damping coefficients, respectively [7]. For every sequence $\mathbf{z}(t)$ of tract variable vectors which solves the differential equation, we have infinitely many sequences of articulatory parameter vectors which describe the mid-sagittal vocal-tract shape. We find one such path $\mathbf{w}(t)$ by solving the differential equation $\ddot{\mathbf{w}} = J^*(-BJ\dot{\mathbf{w}} - K(\mathbf{G}(\mathbf{w}) - \mathbf{z}_0))J^*\dot{J}\dot{\mathbf{w}}$. This follows from the change of variables $\mathbf{z} = \mathbf{G}(\mathbf{w})$, $\dot{\mathbf{z}} = J\mathbf{w}$, $\ddot{\mathbf{z}} = J\dot{\mathbf{w}} + \dot{J}\mathbf{w}$ and from the pseudoinverse J^* of J from [7]. By consecutively solving two such systems, setting the initial conditions of the second to the target articulatory configuration of the first, we can animate the articulatory model to produce VCV sequences.

The first system in the VCV sequence begins with a vocal-tract configuration for a target V1 obtained from rtMRI data by the speaker-specific articulatory model. Specifically, this configuration is found as the average of the weights obtained from all utterances of the given vowel in the MOCHA-TIMIT dataset. The dynamical system then deforms that configuration to achieve a specified constriction for the consonant C, which is chosen from bilabial, alveolar, or velar. The final consonantal configuration is then converted from its representation in tract

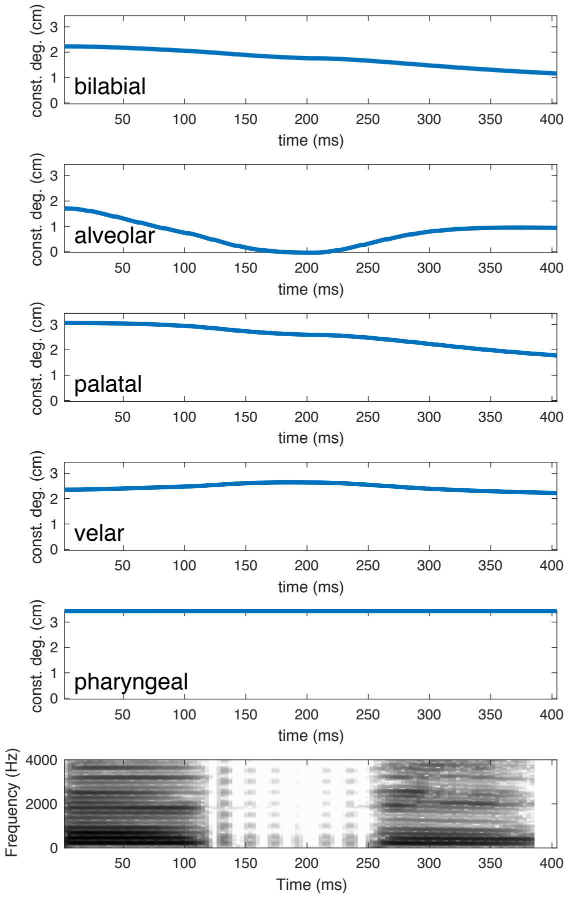

Figure 3: *Tract variables (degrees of constriction), for sequence /adu/ and spectrogram of synthesized sequence.*

variables to that of articulatory parameters, which are used as the initial configuration to the second dynamical system. This second system deforms the constricted vocal tract into the configuration for the vowel V2 (again found as the average in the dataset), whose articulatory parameters are converted into tract variables. The output of both dynamical systems is a matrix of articulatory weights with respect to time. This matrix represents the changing vocal-tract shape throughout the VCV sequence as a sum of the deformations of each articulatory component, and can then be used as the basis of input to the MATLAB synthesizer by specifying the area of the midsagittal slice at each point in time.

2.4. Area Functions and Speech Acoustics

The articulatory parameter trajectories obtained from the dynamical system are then converted to area-function dynamics and input to an articulatory synthesizer based on the simulation developed by Maeda [14, 21]. The dynamic vocal-tract shape is reconstructed from the articulatory parameter specifications and converted to a dynamic area function by superimposing gridlines on the midsagittal slice which segment the vocal tract into 27 sections. The length x and cross-sectional area A of each section are calculated where x is the distance between gridlines and $A = \alpha x^{\beta}$ for a specified value of α and β [22, 23, 24, 25].

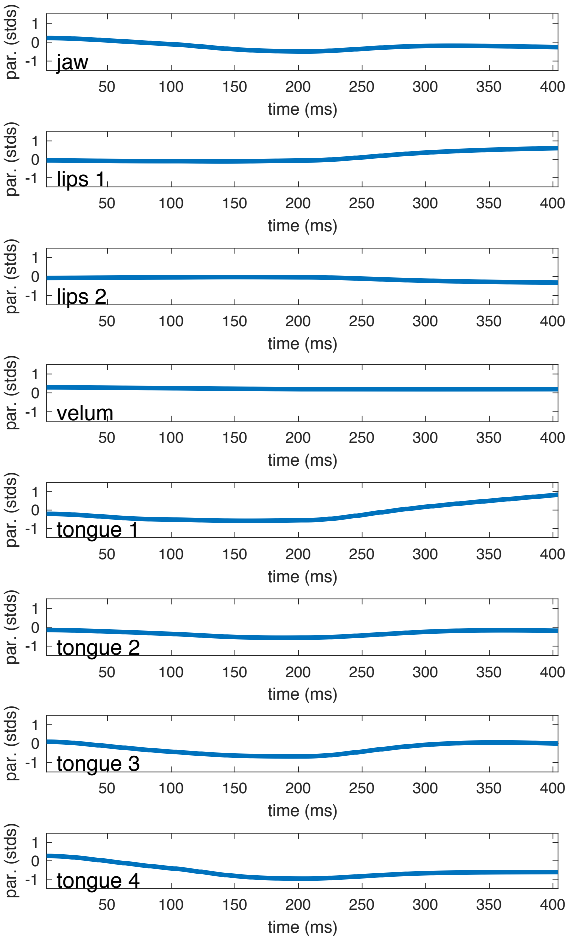

Figure 4: *Parameters of the articulatory model for sequence /adu/.*

This information is converted into 27-dimensional arrays for A and x that describe the area of the vocal tract throughout the VCV sequence.

While area functions of the vocal tract are calculated directly from the dynamical system, controls describing the glottal opening are developed empirically [15, 26]. Inputs to Maeda's synthesizer describing the glottis consist of a slow-varying, non-vibrating component A_{g0} added to a fast-varying triangular glottal pulse with fundamental frequency F0 and amplitude A_p [27, 15]. The use of exclusively voiced consonants in this work allow for A_p to be held constant at 0.2 cm^2, with a smooth cosine transition to and from 0 at the beginning and end of the sequence. Likewise, A_{g0} is held constant at 0 cm^2, as there were no voiceless sounds. The F0 trajectory is generated based on previously recorded VCV utterances from the speaker, and could be further refined to fit the data provided by the dynamical system. As nasal consonants were not synthesized in this work, the area of the velopharyngeal port coupling the nasal and oral cavities is set to 0, and the shape of the nasal cavity is not considered.

The glottal specifications and area functions are input to the synthesizer, which calculates the propagation of sound in a time-varying lumped electrical transmission-line network spec-

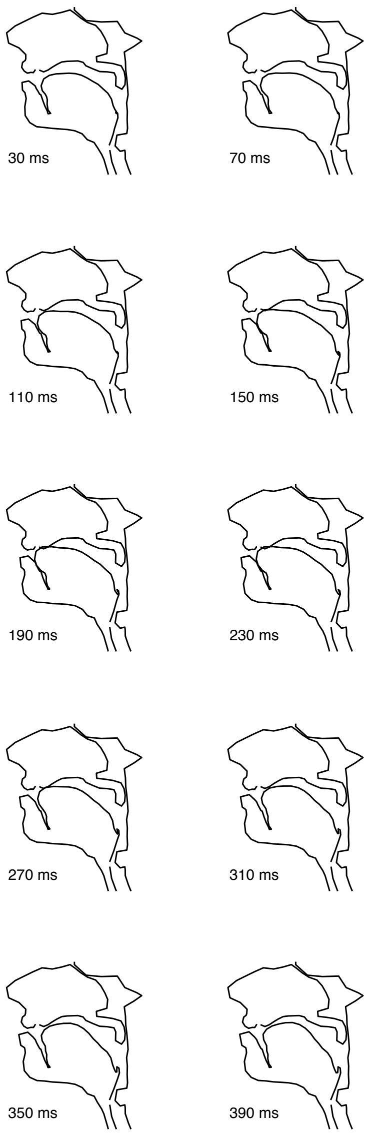

Figure 5: *Temporal evolution of the mid-sagittal vocal-tract shape for the sequence /adu/.*

ified by the given area function dynamics. The acoustic equations through which the glottal signal is passed simulate a system which can be solved at any point in space and time with a backward substitution and elimination procedure, calculating the pressure and volume velocity at each section of the vocal tract. The volume velocity at the lips is differentiated along time to provide the final speech signal.

3. Results

We animated VCV sequences with combinations of 3 vowels (/a/, /i/, /u/) and 3 voiced plosive consonants (/b/, /d, /g/). The constrictions achieved through the dynamical system provide inputs to the synthesizer that successfully produce audible stops, with the area of the vocal tract reaching 0 cm^2 at the appropriate place of articulation. As an illustrative example, for the VCV /adu/, the constriction degree of the alveolar tract variable reached 0 cm (Fig. 3) for about 70 ms, corresponding to a fairly natural-sounding voiced alveolar closure. As expected, the other tract variables displayed no significant closures. The generated result from the synthesizer is illustrated in the spectrogram in Fig. 3, with appropriate voicing and F0 trajectories.

Fig. 4 shows the deformation of the articulatory components used to create this sequence, represented by trajectories of articulatory weights that are converted from the appropriate tract variables, and illustrating that the tongue and jaw are the main factors in creating and releasing the alveolar closure. The dynamics of these components can be used to reconstruct and reanimate the midsagittal slice as it changes over the course of the VCV, as shown in Fig. 5. The vocal tract begins at an open /a/ configuration, and by 170 ms the tongue and jaw have achieved a full alveolar closure.

No formal perceptual evaluation was conducted, but subjectively the utterances sounded appropriate; an alveolar closure generated by the synthesizer was clearly distinguishable from a bilabial or velar one. Most distinctively, the articulatory model and dynamical system were able to accurately animate a closure that was consistent throughout the synthesis architecture; from the manipulation of the tract variables in Fig. 3, to the visual animation in Fig. 5, and finally to the audible results generated by the synthesizer: creating a constriction that successfully generated VCV acoustics with the synthesizer was a primary goal of expanding and combining both frameworks. Further examples can be found online (see Section 1).

4. Conclusion

We have presented an initial architecture for synthesizing VCV sequences based on inputs to a dynamical system that are derived from a factor-based articulatory model and deformed to achieve a target consonantal constriction. Artificially generating these constrictions from rtMRI data allows for inputs to Maeda's synthesizer that achieve precise vocal-tract closures while preserving naturalness.

The combination of these systems provides a promising method for articulatory synthesis and can be expanded in the future to include a more extensive set of VCVs. Voiceless consonants can be generated given further refinement of the temporal coordination between the articulation of the consonant and the opening of the glottis, while nasal consonants can be achieved given a specification of the nasal cavity shape and a similar temporal coordination with the opening of the velopharyngeal port.

The experiments presented here are speaker-specific. We have shown previously that factor-based articulatory models, maps between tract variables and model control parameters, and dynamical systems built with the methods proposed can capture speaker-specific articulatory strategies [13]. It will be interesting to explore to what extent synthesis results will also be different as a function of the modeled speaker.

It may be tempting to think of the methodology we propose here as a building block for a future fully-fledged articulatory text-to-speech synthesis system. For the time being, we think more of its potential use as a tool for conducting analysis-by-synthesis experiments under precise control of articulatory dynamics and coordination, with a view to helping illuminate speech production-perception links [28, 29].

5. Acknowledgments

Work supported by NIH grant R01DC007124 and NSF grant 1514544.

6. References

[1] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: A position paper," in *4th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW-4)*, Perthshire, Scotland, 2001.

[2] Y. Laprie, M. Loosvelt, S. Maeda, R. Sock, and F. Hirsch, "Articulatory copy synthesis from cine X-ray films," in *Interspeech*, Lyon, France, 2013, pp. 2024–2028.

[3] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, p. e60603, 2013.

[4] A. Toutios and S. S. Narayanan, "Articulatory synthesis of French connected speech from EMA data," in *Interspeech*, Lyon, France, 2013, pp. 2738–2742.

[5] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004.

[6] C. P. Browman and L. Goldstein, "Articulatory Phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[7] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.

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

[9] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *The Journal of the Acoustical Society of America*, vol. 70, no. 2, pp. 321–328, 1981.

[10] A. Toutios and S. Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, p. e6, 2016.

[11] S. G. Lingala, A. Toutios, J. Töger, Y. Lim, Y. Zhu, Y.-C. Kim, C. Vaz, S. Narayanan, and K. Nayak, "State-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function," in *Interspeech*, San Francisco, CA, 2016, pp. 475–479.

[12] A. Toutios and S. S. Narayanan, "Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data," in *International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, Aug. 2015.

[13] T. Sorensen, A. Toutios, L. Goldstein, and S. Narayanan, "Characterizing vocal tract dynamics across speakers using real-time MRI," in *Interspeech*, San Francisco, CA, 2016.

[14] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, no. 3-4, pp. 199–229, 1982.

[15] ——, "Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer," in *Sound Patterns of Connected Speech: Description, Models and Explanation*, A. Simpson and M. Pätzold, Eds., 1996, pp. 145–164.

[16] A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," vol. 5, pp. 1–13, 2000.

[17] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

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

[19] A. Lammert, L. Goldstein, S. Narayanan, and K. Iskarous, "Statistical methods for estimation of direct and differential kinematics of the vocal tract," *Speech Communication*, vol. 55, no. 1, pp. 147–161, 2013.

[20] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[21] A. Toutios, T. Sorensen, K. Somandapelli, R. Alexander, and S. S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data," in *Interspeech*, San Francisco, CA, 2016, pp. 1492–1496.

[22] S. Maeda, "Un modèle articulatoire de la langue avec des composantes linéaires," in *Actes 10èmes Journées d'Étude sur la Parole*, Grenoble, France, 1979, pp. 152–162.

[23] ——, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, W. Hardcastle and A. Marchal, Eds. Amsterdam: Kluwer Academic Publisher, 1990, pp. 131–149.

[24] P. Perrier, L.-J. Boë, and R. Sock, "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: modeling the transition with two sets of coefficients," *Journal of Speech, Language, and Hearing Research*, vol. 35, no. 1, pp. 53–67, 1992.

[25] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication*, vol. 36, no. 3, pp. 169–180, 2002.

[26] A. Toutios and S. Maeda, "Articulatory VCV Synthesis from EMA data," in *Interspeech*, Portland, Oregon, 2012, pp. 2566–2569.

[27] G. Fant, "Vocal source analysis – a progress report," *STL-QPSR (Speech Transmission Laboratory, KTH, Stockholm, Sweden)*, vol. 20, no. 3-4, pp. 31–53, 1979.

[28] F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some experiments on the perception of synthetic speech sounds," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 597–606, 1952.

[29] J. Vaissiere, "Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages," in *Experimental Approaches to Phonology*, M. Solé, P. Beddor, and M. Ohala, Eds. Oxford: OUP, 2007, pp. 54–71.