*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# Articulatory Synthesis based on Real-Time Magnetic Resonance Imaging Data

*Asterios Toutios, Tanner Sorensen, Krishna Somandepalli, Rachel Alexander, Shrikanth Narayanan*

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

## Abstract

This paper presents a methodology for articulatory synthesis of running speech in American English driven by real-time magnetic resonance imaging (rtMRI) mid-sagittal vocal-tract data. At the core of the methodology is a time-domain simulation of the propagation of sound in the vocal tract developed previously by Maeda. The first step of the methodology is the automatic derivation of air-tissue boundaries from the rtMRI data. These articulatory outlines are then modified in a systematic way in order to introduce additional precision in the formation of consonantal vocal-tract constrictions. Other elements of the methodology include a previously reported set of empirical rules for setting the time-varying characteristics of the glottis and the velopharyngeal port, and a revised sagittal-to-area conversion. Results are promising towards the development of a full-fledged text-to-speech synthesis system leveraging directly observed vocal-tract dynamics.

**Index Terms**: speech production, articulation, vocal-tract imaging, speech synthesis

## 1. Introduction

The term articulatory speech synthesis is currently used to describe two, quite distinct, families of methods. The first one attempts to synthesize speech by simulating the dynamics of the vocal tract and the propagation of sound therein [1, 2]. The second one uses machine learning-based articulatory-to-acoustic mappings to drive parametric synthesizers [3, 4]. We may refer to the first family of methods as *model-based* articulatory synthesis and to the second one as *machine learning-based*. The present paper considers model-based articulatory synthesis.

Real-time magnetic resonance imaging (rtMRI) has enabled the acquisition of high-speed mid-sagittal imaging data from the entire vocal tract in unprecedented volumes [5, 6, 7], creating new opportunities for addressing problems in speech analysis and technology, such as model-based articulatory speech synthesis, which has classically employed X-ray vocal-tract data [8, 9]. We are working towards building an articulatory text-to-speech architecture, mirroring the TADA effort by Haskins Laboratories [10], directly informed by rtMRI data.

The present paper describes work in this direction. The current system uses an articulatory model, derived from rtMRI data, that represents compactly vocal-tract shapes as weighted sums of articulatory components [11]. While vowels are represented by arrays of articulatory weights directly observed in rtMRI, vocal-tract shapes for consonants are modified from their observed versions via a dynamical system-based model [12] to impose very precise airway constrictions.

The system comprises a sequence of modules. It uses at its core the lossy time-domain simulation of the propagation of sound in the vocal tract proposed by Maeda [13, 14], but this is not restrictive. It is perfectly imaginable that elements of the present work can be used in tandem with other models.

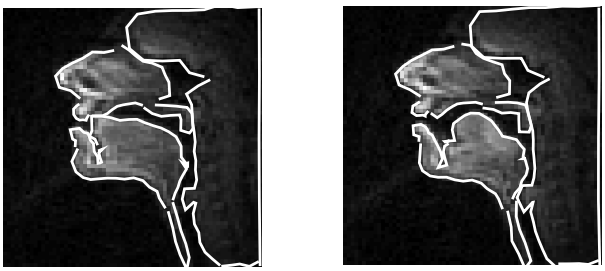

Figure 1: *Demonstration of results of automatic air-tissue boundary segmentation algorithm. Left and right frames show respectively /s/ and /r/ frames from the utterance "This was easy for us".*

## 2. Method

### 2.1. Automatic derivation of air-tissue boundaries

We used rtMRI data from the F1 speaker of the USC-TIMIT database [5], a 23-year old female speaker born in New York. The speech material recorded with rtMRI in the particular session corresponded to the 460 sentences of the MOCHA-TIMIT dataset [15]. The data comprise videos of mid-sagittal vocal-tract images at 23.18 frames per second. The size of the images is $68 \times 68$ pixels and the spatial resolution is 3 millimeters per pixel. Synchronized audio data, recorded concurrently with rtMRI, and carefully prepared transcriptions, were used to generate aligned labels at the phonetic level with the freely available tool SailAlign [16].

The video data were subjected to an updated version of an automatic segmentation algorithm previously published [17, 11]. The segmentation method considers the outlines of 15 anatomical features comprising three connected regions of tissue. For every image in a video sequence, the method incrementally deforms an initial set of anatomical feature outlines (a template), by displacing the fixed number of points comprising each outline, until a fit to the observed image data is achieved. After this process, there are 184 points in total describing the articulatory contours corresponding to each rtMRI frame. Fig. 1 shows two typical examples of segmentation results.

### 2.2. Model-based refinement of constriction degrees

The outlines derived by the above procedure should be considered accurate enough for articulatory synthesis in the case of vowels. However, this may not be so for fricative and stop consonants, because of the spatial resolution of the rtMRI data



http://dx.doi.org/10.21437/Interspeech.2016-596

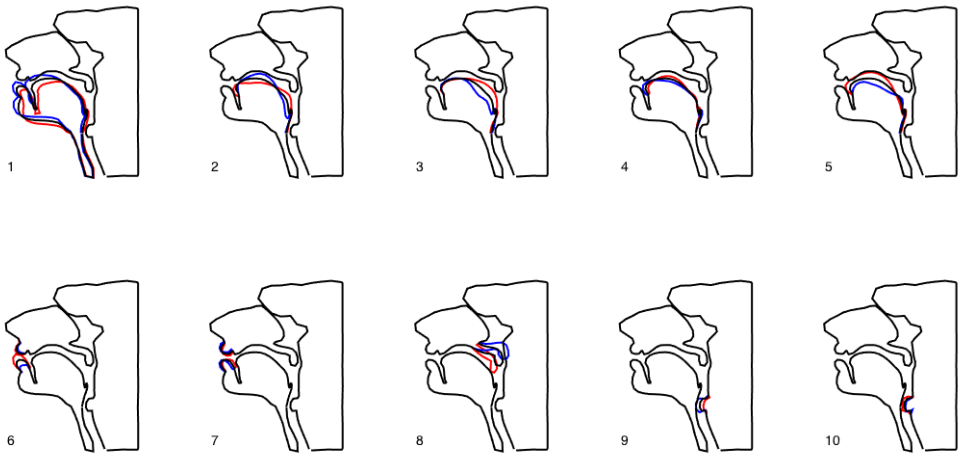


Figure 2: *Components of the articulatory model used in this study. Colored line segments denote mean and ±2 standard deviations.*

and potential noise introduced by the automatic segmentation method. From previous experiments [18, 19] we have seen that the cross-sectional area of the vocal-tract at the place of articulation should optimally be zero during the closure phase of stops and 0.1 $cm^2$ during fricatives. The latter translates to a sagittal distance of ∼3.5 mm between the active and passive articulator (constriction degree) assuming a cylindrical shape for the vocal-tract section. This is comparable to the rtMRI spatial resolution of 3 mm/pixel.

In those previous experiments we had enforced the area function to assume these values *a posteriori*, i.e., after its derivation from sagittal slices and just before using it as input to the synthesizer (see later sections of this paper for the full process). Though this operation had given good synthesis results, it could not be assumed that the resulting area functions would correspond to vocal-tract shapes which could be produced naturally by the speaker. In this section we describe a process that introduces corrections of constrictions degrees in the sagittal slices, in a way that preserves their naturalness.

Recently [11], we described a factor analysis method for deriving a mid-sagittal articulatory model, and applied it to data from the same speaker, and recording session, to the one used in the present paper. The model decomposes the mid-sagittal slice into a weighted sum of articulatory components (Fig. 2). Since the components are fixed throughout the dataset, the mid-sagittal slice dynamics are well approximated by trajectories of the applied weights (or, articulatory parameters). We have also introduced [12] a locally linear mapping, based on an hierarchical clustering process, from such articulatory parameters to constrictions between passive and active articulators along the vocal tract. This was used to build a dynamical system that finds, based on concepts from Task Dynamics [20], an optimal (in a least-squares sense) path from an initial mid-sagittal vocal-tract configuration (array of articulatory model parameters) to a configuration that achieves a particular constriction (combination of place of articulation and constriction degree, see Fig. 3).

For the present work, based the phonetic alignments of the synchronized audio files, we identify all rtMRI frames corresponding to the closure phase of stops (including nasal stops), from beginning to burst (which we locate at at 70% of the stop duration) and the full duration of fricatives. For each of these frames, we run the dynamical system targeting the desired constriction (combination of place of articulation and constriction degree) using the articulatory parameters of the frame as the initial vocal-tract configuration. The result, for each frame, is a corrected array of articulatory parameters that: (i) achieves the desired constriction; (ii) is minimally different from the segmentation result for the particular frame; and (iii) can be expressed as a weighted sum of the factors of the articulatory model, which indicates that it can be naturally produced by the speaker. Note that, because constrictions are achieved by synergies of articulatory components [21, 22], the corrected configuration differs from the original in multiple articulatory parameters.

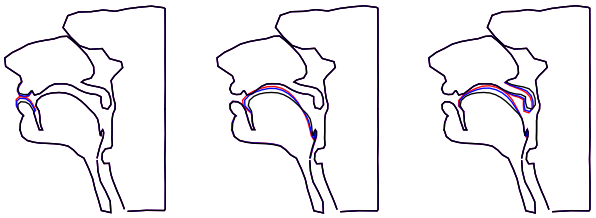

Figure 3: *Illustration of the dynamical system used in this study. Red and blue lines represent the time-course of vocal-tract movements that optimally achieve a (from left to right) bilabial, alveolar, or velar closure beginning from a neutral posture (black line). (The three lines represent shapes that are equidistant in time.)*

Fig. 4 shows the articulatory parameter trajectories for one utterance, both original and those corrected by the above operation. Sagittal-slice dynamics are reconstructed from the corrected trajectories and are used as input to the next module of the synthesis system (Fig. 5).

### 2.3. Sagittal-to-area conversion

A scaled version of the semi-polar grid proposed by Maeda [23, 24] is superposed on the mid-sagittal contours that have been reconstructed from the modified articulatory parameters, as shown in Fig. 6, and the intersections of contours and gridlines (starting at the glottis) are found. This information, supplemented by measurements of the lip opening (minimum distance between the upper and lower lip contours) and lip protrusion

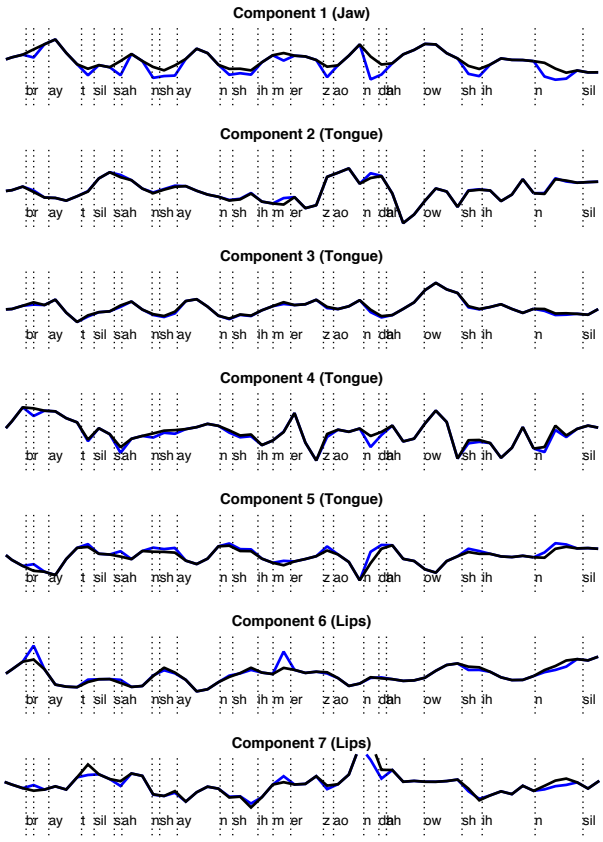


Figure 4: *Articulatory parameter trajectories for the utterance "Bright sunshine shimmers on the ocean". Original trajectories are shown in black. Blue line segments show the corrections applied using the dynamical system discussed.*

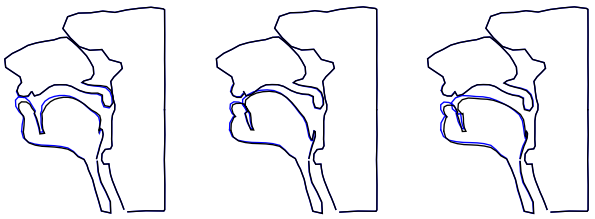

Figure 5: *Vocal-tract sagittal slices reconstructed from articulatory parameters of Fig. 4 for consonants /b/, /t/, and /s/ in "**b**righ**t** **s**unshine..." (black: before correction; blue: after correction)*

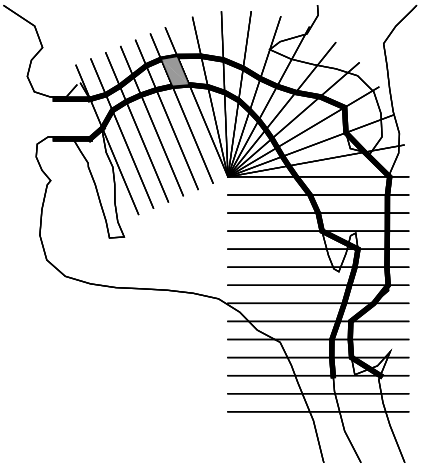

Figure 6: *Semi-polar articulatory grid superposed on a mid-sagittal slice. Intersections of the grid with the slice, supplemented by minimal lip opening information, define the internal and external tract walls (thicker lines). The conversion to area function is based on trapezoids like the one shown shaded (see text).*

(horizontal distance of the point of minimal lip separation from the first grid-line), defines the inner and outer vocal-tract walls.

The distance $d$ between the points of intersection of any gridline and the vocal-tract walls, and the vocal-tract cross sectional area $A$ at that gridline, are related by $A = \alpha d^{\beta}$, where $\alpha$ and $\beta$ are different for each gridline, and have been given by Maeda (see also [25, 26]). Given this information, the volume $V$ of a 3-dimensional polygon corresponding to any trapezoid defined by the intersections of the two vocal-tract walls and two adjacent gridlines (such as the shaded trapezoid in Fig. 6), can be calculated. Each 3D polygon is then converted to a cylinder with same volume $V$ and length $x$ equal to the distance $x$ between the midpoints of the two "gridline" sides of the polygon. The areas of the circular faces of the cylinders, together with their lengths, define the area function corresponding to the sagittal slice. Overall, area functions are described by 27-dimensional arrays of values for $A'$ and $x$.

### 2.4. Glottis, nasal port, F0 and acoustic simulation

Maeda's simulator uses a modified version of a model for the glottis proposed by Fant [27]. Glottal area is modeled as the sum of a slow and a fast-varying component [14]. The fast-varying component is a triangular glottal pulse with amplitude $A_p$ and fundamental frequency F0 which is added to a non-vibrating (slow-varying) area component $A_{g0}$. In our experiment, $A_p$, and $A_{g0}$ were set in a fashion similar to [19]: throughout the duration of voiced segments (as identified in the phonetic alignments) $A_p$ was set to 0.2 cm$^2$ and $A_{g0}$ to zero; in unvoiced segments $A_p = 0$ and $A_{g0} = 0.4$ cm$^2$. Between these two extreme cases, values of the two glottal components varied smoothly by a raised cosine transition. The area of the velopharyngeal port that couples nasal and oral cavities was set to 0.4 cm$^2$ during nasal segments and zero elsewhere. The nasal area function provided with Maeda's simulator was used without any adaptation to our speaker. Note that the controls for glottis and the velopharyngeal port could potentially be inferred from the sagittal slices, i.e., from information on the shaping of the arytenoid and the velum, respectively, without resorting to the phonetic labels. This is something we will further explore in the future.

F0 is set to follow a linear trajectory from 250 Hz at the beginning of any utterance to 160 Hz at its end. In previous synthesis experiments with EMA [19] we had used F0 trajectories extracted from the concurrently recorded audio information. However, because of noise in the rtMRI audio data (even after de-noising to account for the very loud scanner noise, there is still residual noise), automatic F0 tracking did not give vey

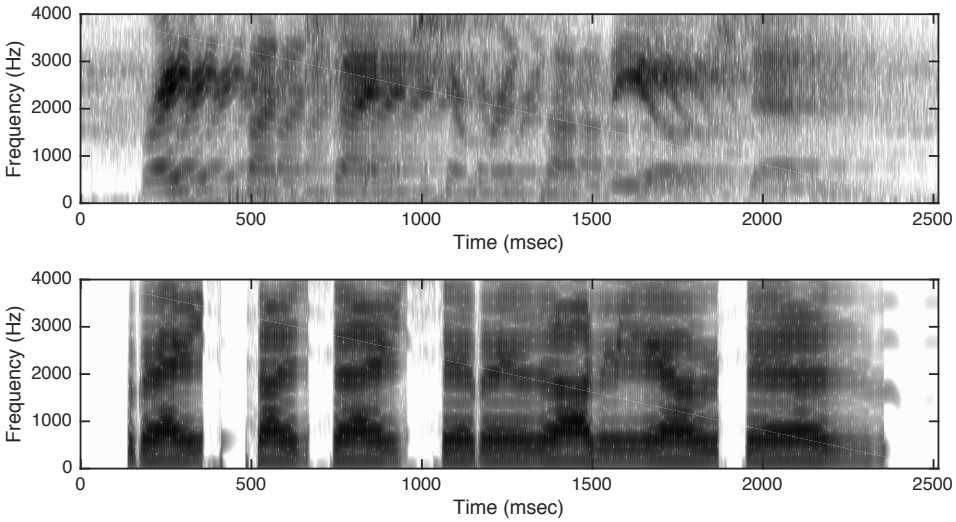


Figure 7: *Spectrograms of the audio recorded concurrently with rtMRI (top), and the synthesized audio (bottom) for the utterance "Bright sunshine shimmers on the ocean". Note that the recorded audio is noisy (even after de-noising to account for very loud scanner noise, there is still residual noise).*

useful results.

All the above controls, including area functions, are used as input to Maeda's simulator [13], which represents the vocal tract as a lumped electrical network and calculates the propagation of sound therein. Friction noise is generated automatically in the model, when correct aerodynamic conditions are met [14].

## 3. Results

We synthesized the 20 first sentences of the MOCHA-TIMIT set; these synthesis results are available at `http://sail.usc.edu/span/rtmri_synthesis/`. Fig. 7 compares the spectrograms of the audio recorded concurrently with rtMRI and the synthesis result for one of the utterances. Though we cannot make hard claims about the intelligibility and naturalness of the synthesis results, most of the utterances sounded fairly intelligible. Particularly, we were able to accurately model the constrictions for consonants which was the main goal of the work presented herein. With the exception of a few artifacts, segments sounded well co-articulated. Naturalness can be further improved by applying more realistic F0 trajectories.

We have attempted an objective evaluation by calculating the Normalized Root Mean Squared Error (NRMSE) between Mel Frequency Cepstral Coefficients extracted from the recorded and synthesized audio. However, the results were not at par with previous reports in related literature. It may be that noise in the recorded data skews the NRMSE calculation.

The MOCHA-TIMIT sentences, because of their peculiar semantics, may not be optimal for assessing intelligibility. To best assess intelligibility, we plan to design an rtMRI data collection of minimal pair words and sentences, and examine if, after synthesis, such minimal pairs can be perceptually differentiated

## 4. Concluding Remarks

We have presented work on model-based articulatory speech synthesis based on rtMRI speech production information. The current system can be considered as a hybrid one, since information for vowels is derived directly from rtMRI observations (after automatic delination of air-tissue boundaries, and a compact representation thereof), while vocal-tract controls for consonants are derived by modifying observed slices according to a dynamical system, in order to correctly account for vocal-tract constrictions. The modules of the system are rather independent, and alternatives may be considered. For example the method described in [28] to find the centerline of the vocal-tract might lead to more precise area functions than the application of the articulatory grid. Refining the estimates of the $\alpha$ and $\beta$ coefficients in the sagittal-to-area conversion using data from an accelerated volumetric MRI protocol [29] may be beneficial as well. We are also developing a Matlab version of Maeda's simulator, which will help make our overall architecture even more transparent and consistent (all other modules are already implemented in Matlab).

The spatiotemporal resolution (3 mm/pixel, 23.18 frames per second) of the rtMRI data we used may not be optimal. We plan to repeat our experiments using data collected with a recently developed protocol at higher resolution (2.4 mm/pixel, 83 frames per second [30]), which will provide more accurate sagittal dynamics.

Ultimately, we want to build an articulatory synthesis system that does not use directly observed data at runtime, but rather generates articulatory controls from text. We are considering specifying targets for vowels in the space of the weights of the articulatory model, or even in the formant space (and derive the articulatory weights by inversion [31]), while consonants will be specified in the vocal-tract constriction space. Such a system should also be combined with a prosodic model to derive relative timings of syllables [32, 33]. These efforts will be informed and guided by speech production data, such as real-time MRI.

## 5. Acknowledgment

Work supported by NIH grant R01DC007124 and NSF grant 1514544.

# 6. References


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

[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.

[3] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[4] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016.

[5] 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, 2014.

[6] J. Kim, A. Toutios, Y.-C. Kim, Y. Zhu, S. Lee, and S. S. Narayanan, "USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging," in *International Seminar on Speech Production (ISSP)*, Cologne, Germany, May 2014.

[7] 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.

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

[9] A. Bothorel, P. Simon, F. Wioland, and J. Zerling, *Cinéradiographie des voyelles et consonnes du français*. L'Institut de Phonétique de Strasbourg, France, 1986.

[10] 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.

[11] 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.

[12] T. Sorensen, A. Toutios, L. Goldstein, and S. S. Narayanan, "Characterizing vocal tract dynamics with real-time MRI," in *15th Conference on Laboratory Phonology*, Ithaca, NY, Jul. 2016.

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

[14] ——, "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.

[15] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *5th Seminar on Speech Production*, Kloster Seeon, Germany, 2000, pp. 305–308.

[16] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA, 2011.

[17] 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, Mar. 2009.

[18] A. Toutios and S. Maeda, "Articulatory VCV Synthesis from EMA data," in *Interspeech*, Portland, Oregon, 2012.

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

[20] 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.

[21] C. A. Fowler, P. Rubin, R. E. Remez, and M. E. Turvey, "Implications for speech production of a general theory of action," 1980.

[22] M. L. Latash, *Synergy*. Oxford University Press, 2008.

[23] 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.

[24] ——, "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.

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

[26] 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.

[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] S. Maeda and Y. Laprie, "Vowel and prosodic factor dependent variations of vocal-tract length," in *Interspeech*, Lyon, France, 2013.

[29] Y.-C. Kim, S. S. Narayanan, and K. S. Nayak, "Accelerated three-dimensional upper airway MRI using compressed sensing," *Magnetic Resonance in Medicine*, vol. 61, no. 6, pp. 1434–1440, 2009.

[30] S. G. Lingala, Y. Zhu, Y.-C. Kim, A. Toutios, S. Narayanan, and K. S. Nayak, "A fast and flexible MRI system for the study of dynamic vocal tract shaping," *Magnetic Resonance in Medicine*, 2016.

[31] 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.

[32] F. Bell-Berti and K. S. Harris, "A temporal model of speech production," *Phonetica*, vol. 38, no. 1-3, pp. 9–20, 1981.

[33] O. Fujimura, "The C/D model and prosodic control of articulatory behavior," *Phonetica*, vol. 57, no. 2-4, pp. 128–138, 2000.