# Articulatory Synthesis of French Connected Speech from EMA Data

*Asterios Toutios, Shrikanth S. Narayanan*

Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, USA

{toutios,shri}@sipi.usc.edu

## Abstract

This paper reports an experiment in synthesizing French connected speech using Maeda's digital simulation of the vocal-tract system. The dynamics of the vocal-tract shape are estimated from the dynamics of Electromagnetic Articulograph (EMA) sensors via Maeda's geometrical articulatory model. Time-varying characteristics of the glottis and the velopharyngeal port are set using empirical rules, while the fundamental frequency pattern is copied from the concurrently recorded audio signal. A subjective experiment was performed online to assess the perceived intelligibility and naturalness of the synthesized speech. Results indicate that a properly driven simulation of the vocal tract has the potential to provide a scientifically grounded alternative to the development of text-to-speech synthesis systems.

**Index Terms**: articulatory synthesis, speech production, electromagnetic articulography, vocal-tract simulation

## 1. Introduction

An ultimate test for a model of the relationship between the vocal-tract configuration and the speech waveform, would be to use it to generate highly intelligible and natural-sounding speech. If such a test were successful, it would open up new scientifically grounded avenues to the development of text-to-speech synthesis systems [1]. It would also allow for a veritable means for studying details of articulatory-acoustic relations, and broader questions pertaining to production-perception links [2].

Models generating speech waveforms from vocal-tract configurations, i.e. articulatory synthesizers or acoustic-to-articulatory simulations, have been in place for several years now [3, 4]. However, they are most often used to study the articulatory-to-acoustic relationships for isolated static configurations of the vocal tract, and for continuant sounds like vowels or fricatives. In broad terms, the vocal-tract configurations for such sounds are part of standard phonetic knowledge, at least for major languages.

On the other hand, the potential of using such models to synthesize connected speech has been much less addressed. It is well-known that connected speech cannot be considered merely as the result of a straightforward concatenation of static speech sounds, neither at the acoustic nor at the articulatory level [5]. To synthesize connected speech, one needs to provide as input to the articulatory-to-acoustic simulation the dynamics of the vocal-tract configuration.

Perhaps the best-known method to generate such inputs is the Task Dynamics Model [6], which calculates the control parameters of Rubin et al.'s articulatory synthesizer [3], based on the theory of Articulatory Phonology [7]. However there is no explicit guarantee that the articulatory control trajectories thus derived correspond to the actual articulations of any given speaker, which raises an additional complication: assuming imperfections in the generated speech waveforms, one cannot tell if they are due to problems in the Task Dynamics Model, or due to problems in the articulatory-to-acoustic simulation itself.

This paper describes an experiment that follows a different approach to the problem of driving an articulatory synthesizer, specifically Maeda's synthesizer [4]. We use Electromagnetic Articulography (EMA) data as the primary input source and attempt to re-synthesize the concurrently recorded speech waveform. The dynamics of the vocal-tract shape (i.e. jaw, lips and tongue), as represented in the context of Maeda's articulatory model, are extrapolated non-trivially on the basis of the EMA dynamics. These dynamics are augmented by characteristics of the glottal area and nasal coupling, which are determined by simple rules on the basis of phonetic content. To avoid possible problems due to the interplay between vocal-tract resonance frequencies and the fundamental frequency of glottis vibration, we copy the F0 trajectory from the recorded speech waveform. This approach may enable us to focus on imperfections of the core function of the articulatory synthesizer by decoupling it from the task of artificially generating plausible articulatory control inputs.

In previous work [8] we have attempted to synthesize vowel-consonant-vowel sequences, where the consonant was an unvoiced oral stop or fricative, using a similar method. The present paper takes this idea one step further and applies it to the synthesis of complete utterances including voiced and nasal consonants. For unvoiced consonants, the rules for controlling glottal area characteristics have been revised and simplified.

## 2. From EMA to vocal-tract shapes

A method to estimate the control parameters of Maeda's articulatory model from EMA data has been previously presented by Toutios at al. [9]. The present work builds upon that contribution, and reports an extended and updated version of the approach. The description begins with a brief introduction to Maeda's geometrical articulatory model.

### 2.1. Maeda's articulatory model

Maeda's geometrical articulatory model [10, 11] describes the (oral part of the) vocal-tract shape by means of seven parameters(Fig. 1a): jaw opening ($p_1$); tongue dorsum position ($p_2$); tongue dorsum shape ($p_3$); tongue apex position ($p_4$); lip opening ($p_5$); lip protrusion ($p_6$); and larynx height ($p_7$). The 7-tuples of articulatory parameters specify mid-sagittal profiles of the vocal-tract, plotted over a pre-defined semi-polar grid(Fig. 1b). More specifically, measurements on the articulatory grid and the lips, collectively called *variables*, can be derived as linear combinations of control parameters.

The *tongue variables* are defined as the coordinates of the intersections of the tongue contour with the grid. Out of the 31

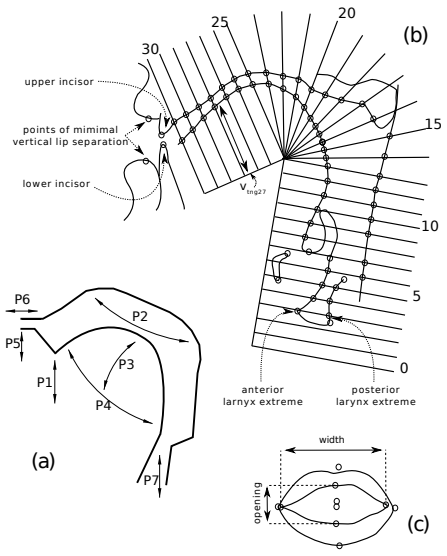


Figure 1: *Maeda's articulatory model: (a) control parameters; (b) articulatory grid; (c) definitions of lip variables.*

grid-lines, numbered as shown in Fig. 1b, the 6th to 30th are related to the tongue contour. Thus, variables $v_{tng6}, \ldots, v_{tng30}$ are defined. In Fig. 1b, the measurement of $v_{tng27}$ is shown as an example. The variables describing the form of the larynx are the $x, y$ coordinates of the anterior ($v_{\text{a,x}}, v_{\text{a,y}}$) and posterior ($v_{\text{p,x}}, v_{\text{p,y}}$) extremes of the larynx with respect to the linear coordinate system. The *lip opening variable* ($v_{ope}$) is defined as the distance between the highest and lowest points on the front inner lip contours. The *lip width variable* ($v_{wid}$) is defined as the distance between the most left and right points on the same contours. The *lip protrusion variable* ($v_{pro}$) is measured on the lip profile as the distance between the upper incisors and the point of the minimal vertical separation between the upper and lower lips. The *jaw variable* ($v_{jaw}$) is defined as the negative of the distance between the upper and lower incisors, projected on the direction of the lines of the linear region of the grid in the buccal area. All variables are $z$-scored.

According to the definition of the articulatory model, the variables described above are generated, at any given instant, from an underlying set of model parameters via a set of linear relationships. The jaw parameter is equal to the jaw variable ($p_1 = v_{jaw}$),and then the model provides the matrices $\mathbf{A}_{\text{lip}}$, $\mathbf{A}_{\text{tng}}$ and $\mathbf{A}_{\text{lrx}}$, so that:

$$[v_{\text{jaw}}, v_{\text{pro}}, v_{\text{ope}}, v_{\text{wid}}]^T = \mathbf{A}_{\text{lip}}\,[p_1, p_5, p_6]^T \quad (1)$$

$$[v_{\text{jaw}}, v_{tng6}, v_{tng7}, \ldots, v_{tng30}]^T = \mathbf{A}_{\text{tng}}\,[p_1, p_2, p_3, p_4]^T \quad (2)$$

$$[v_{\text{jaw}}, v_{\text{a,x}}, v_{\text{a,y}}, v_{\text{p,x}}, v_{\text{p,y}}]^T = \mathbf{A}_{\text{lrx}}\,[p_1, p_7]^T \quad (3)$$

Our goal is to infer the values of the control parameters from EMA data *on a frame-by-frame basis*, which will then enable the re-construction of the mid-sagittal vocal-tract shape through the above equations. We exclude the larynx parameter, since EMA data (at least with the setup we used) cannot provide information on the larynx position. The rest of the parameters *can* be inferred, by the process we will describe in the following paragraphs.

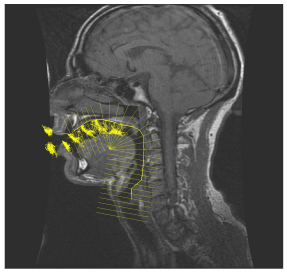

Figure 2: *Registered EMA data and articulatory grid superposed on an MRI image of the speaker*

### 2.2. EMA data, processing and registration

EMA data corresponding to ten short French sentences, known to the speech production community as *the Strasbourg sentences* were recorded at LORIA, Nancy, using the AG500 articulograph [12], as part of a larger EMA dataset. The subject was a phonetically aware native French male speaker. The data comprised three-dimensional dynamics of four sensors attached along the surface of the tongue on the mid-sagittal plane, from the apex to the vicinity of the velum, and sensors on the lower incisor, lower lip, upper lip and lip corners. Additional sensors on the bridge of the nose and behind the ears were used for head movement correction. The sample rate of these data was 200 Hz. The audio signal was recorded simultaneously and synchronized automatically using the articulograph's internal software.

These data were 3-dimensional and we had to project them onto the mid-sagittal plane. The question was then how to define carefully the mid-sagittal plane in the 3D space. We applied Principal Component Analysis on the 3D positions of the sensors on the upper and lower lip, jaw and tongue. We found that more than 99% of the variance in the movement of these sensor lies on a plane, which we considered as our mid-sagittal plane, and on which we projected our sensor data.

The size of the articulatory grid was corrected by adjusting pre-defined *mouth and pharynx scale factors* and its external wall was re-drawn, to account for anatomical differences between the speaker involved in this experiment and the (female) speaker on which the articulatory model was initially constructed.

The next step was to co-register geometrically the EMA data (projected on the mid-sagittal plane) and the articulatory grid. This was done visually, by plotting both over an MRI image of the speaker and making manually the necessary adjustments of coordinate systems (shifts and rotations). The end-result of this process is shown in Fig. 2.

### 2.3. Derivation of jaw and lip parameters

The jaw variable is derived from EMA simply by projecting the jaw sensor position on the left-most line of the articulatory grid. The lip protrusion variable is derived as the distance of the mid-point between upper and lower lip sensors from the left-most grid-line. The lip width variable is the distance between the 3D positions of the sensors on the lip corners (which have not been projected onto the mid-sagittal plane).

It is possible to think that the lip opening variable should be the z-scored distance between the upper and lower lip sensors. However, this would not be correct because of the problem depicted in Fig. 3, which shows classic examples of lip contours

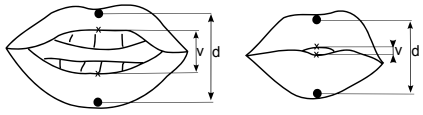


Figure 3: *Typical lip configurations for /i/ (left) and /u/ (right) from [13]. Dots indicate approximate sensor positions.*

for /i/ and /u/ where probable positions of the sensors have been superposed. The distance between the highest and lowest points on the front *inner* lip contours $v$, which is how lip opening is defined in the context of the model, does not correspond to the sensor distance $d$ but is also influenced (apparently) by protrusion.

Though the exact relationship between $v$ and $d$ in Fig. 3 is a matter that calls for further investigation we found that for our purposes (and for our speaker) the following heuristic produced good results: the lip opening variable is derived as the upper and lower lip sensor distance *minus* our protrusion measurement.

The jaw parameter $p_1$ is a z-scored version of the jaw variable. Z-scored versions of the jaw and lip variables described are plugged into Eq. 1 which we solve in the least-squares sense to get parameters $p_5, p_6$ for the frame in question.

### 2.4. Derivation of tongue parameters

At a specific frame of EMA data, let a tongue sensor $i$ be positioned between two grid-lines, of the linear part of the grid, numbered $n$ and $n+1$ at corresponding distances $d_n$ and $d_{n+1}$ from the two grid-lines. Let $\mathbf{a}_n$ and $\mathbf{a}_{n+1}$ be the rows of table $\mathbf{A}_{tng}$ that correspond to these lines. We then derive the vector

$$b_i = \frac{d_{n+1}}{d_n + d_{n+1}}\mathbf{a}_n + \frac{d_n}{d_n + d_{n+1}}\mathbf{a}_{n+1} \tag{4}$$

and assume that

$$v_i = \mathbf{b}_i \left[p_1, p_2, p_3, p_4\right]^T \tag{5}$$

where $v_i$ is the distance of the sensor from the base of the grid-lines (in the same vein as $v_{27}$ shown in Fig. 1). This is equivalent to inserting a *virtual grid-line* at the exact position of the sensor and inferring the linear combination that ties sensor position and parameters at that exact place between tongue variable and tongue parameters for that grid-line.

If the sensor lies in the polar part of the grid, the distances of the above equations are simply replaced by angles. We then have a system of equations like (5) for $i = 1 \ldots 4$, with known $v_i$ which we solve in the least square sense to find the tongue parameters $p_2, p_3, p_4$. An additional constraint we add to this optimization problem is that the tongue contour should not cross the external wall of the vocal tract.

### 2.5. Sagittal vocal-tract profiles and area functions

Having the six parameters, and setting the larynx parameter to its mean (zero) value, we can reconstruct the sagittal vocal-tract profiles for each frame of our data. Three snapshots of the result are shown in Fig. 4.

These sagittal vocal-tract profiles are converted to area functions, consisting of 17 sections of equal length. This length is subject to dynamic change, as a function of vocal-tract shape. The formula $A = \alpha x^{\beta}$ is used, where $A$ and $x$ are respectively the cross-sectional area and the midsagittal vocal-tract opening (distance between internal and external vocal tract) in cm. The

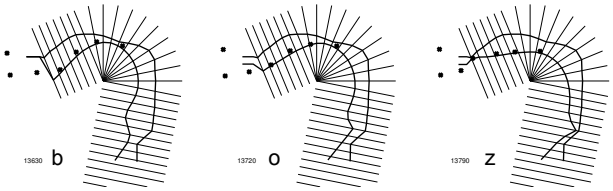


Figure 4: *Snapshots of mid-sagittal vocal-tract slices derived from EMA (registered sensors shown with crosses) taken from the utterance Mets tes* ***beaux*** *habits.*

values of $\alpha$ and $\beta$ vary along the vocal tract, and were determined by Maeda in an *ad hoc* manner [11].

## 3. Further controls for synthesis

Besides the (oral) area function, the dynamics of a few other parameters, related to the status of the glottis and nasal coupling, need to be added as inputs to the articulatory synthesizer. Finally, some semi-automatic corrections are applied to the area functions inferred from EMA, to properly account for consonantal events.

### 3.1. Glottis

Maeda's synthesizer uses a modified version of a model for the glottis proposed by Fant [14]. Glottal area is modeled as the sum of a slow and a fast-varying component [15]. 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 $A_{g0}$.

For setting $A_p$ and $A_{g0}$ in our synthesis experiments, we were inspired by Flanagan, Ishizaka and Shipley [16]. $A_{g0}$ was set to zero during voiced sounds and silent stretches. For unvoiced consonants, starting from zero at the beginning of the consonant (as seen on the spectrogram), $A_{g0}$ reaches a maximum of 0.4 cm$^2$ at 70% of the duration of the vowel, then falls back to zero at the end of the consonant. Between these three extremes, $A_{g0}$ varies smoothly by a raised cosine function.

$A_p$ is set to 0.2 cm$^2$ during voiced sounds and to zero during unvoiced sounds and silent stretches. At the edge between a voiced sound and silence, there is a cosine transition between the two values that spans 20 ms into the silence. At the edge between a voiced and an unvoiced sound, the cosine transition spans 20 ms into the voiced sound. As explained earlier, $F0$ is copied from the concurrent speech recording.

### 3.2. Nasal coupling

For the present synthesis experiment, we considered the nasal area function implemented in Maeda's synthesizer, without adaptation to our speaker. Coupling of nasal and oral cavities is controlled by adjusting the area $A_{nc}$ of the velopharyngeal port. This was set empirically to 0.01 cm$^2$ during oral sounds and silent stretches (setting the value to zero raised computational problems in the simulation), and to 0.4 cm$^2$ during nasals. There was a cosine transition between the two values both at the beginning and ending of the nasal, which spanned 20 ms into the nasal and 30 ms into the adjacent sounds.

### 3.3. Refinement of vocal-tract constrictions

The area of the narrowest constriction of the vocal-tract should be set very precisely to properly distinguish between stops,

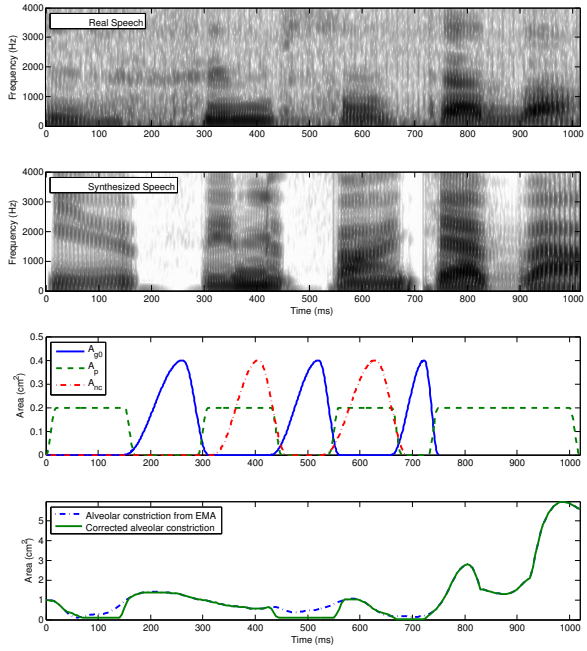


Figure 5: *Synthesis results and (subset of) inputs for the utterance Il fume son tabac. Top to bottom: (i) spectrogram of recorded speech – note that audio is noisy; (ii) spectrogram of synthesized speech; (iii) slow-varying component of glottal area $A_{g0}$; amplitude $A_p$ of fast-varying component of glottal area; area of velopharyngeal port $A_{nc}$ (iv) cross-section area at the alveolar ridge (16th out of 17 cross-sections, numbered from glottis.*

fricatives, and high vowels. We found that the specifications of area function derived directly from EMA were sometimes not precise enough. During stops (including nasal stops), from their beginning to the assumed burst (at 70% of the duration) we forced the area function at the narrowest constriction cross section (or combination of cross-sections) to be equal to 0.001 $cm^2$ (a zero value would raise numerical problems). For the full duration of fricatives, we set the area at the narrowest constriction at 0.1 $cm^2$. For any other cross-sections during those sounds, and for all cross-sections during vowels, we set a low threshold of 0.3 $cm^2$ to ensure an open air passage. When applying these operations, we also replaced area function trajectory values spanning a neighborhood of 30 ms into and out of the forced transitions by interpolated values, to avoid discontinuities of the derivatives of the trajectories that would give rise to audible artifacts (clicks) in our synthesis results.

## 4. Results

We synthesized the ten *Strasbourg sentences* by incorporating the elements described so far to Maeda's synthesizer. In essence, our inputs were (i) the EMA data, (ii) phonetic transcriptions of the sentences, and (iii) the F0 trajectory, as tracked from the audio recordings corresponding to the EMA data. The only change to the original synthesizer of Maeda concerned the generation of friction noise: instead of locally at the narrowest constriction, friction noise was generated along the whole vocal-tract, in the manner and for reasons explained in [8]. Fig. 5 shows spectrograms of the recorded and synthesized speech for a segment of our results,

Table 1: *Mean opinion scores (standard deviations in parentheses) for each sentence in terms of intelligibility and naturalness, across 12 anonymous speakers of French (5 being best, 1 being worst).*

| Sentence | Intelligibility | Naturalness |
|---|---|---|
| Ma chemise est roussie | 2.75 ( 1.06 ) | 3.00 ( 0.74 ) |
| Voilà des bougie | 2.75 ( 1.06 ) | 2.75 ( 0.75 ) |
| Donne un petit coup | 3.42 ( 1.13 ) | 3.33 ( 0.89 ) |
| Une réponse ambiguë | 2.33 ( 1.31 ) | 2.75 ( 1.06 ) |
| Louis pense à ça | 2.16 ( 1.00 ) | 2.50 ( 1.00 ) |
| Mets tes beaux habits | 3.33 ( 1.31 ) | 3.17 ( 0.93 ) |
| Une pâte à choux | 1.83 ( 1.31 ) | 2.67 ( 1.15 ) |
| Prête-lui seize écus | 1.75 ( 1.00 ) | 2.25 ( 1.06 ) |
| Chevalier du gué | 2.41 ( 1.15 ) | 2.92 ( 1.00 ) |
| Il fume son tabac | 3.00 ( 1.04 ) | 3.08 ( 1.16 ) |
| **Overall** | 2.58 ( 1.21 ) | 2.84 ( 1.00 ) |

together with trajectories of some of the elements we have described in the present paper. Audio files of the recorded and synthesized versions of the ten sentences can be found at http://sail.usc.edu/span/frenchsynthesis/.

We performed a subjective evaluation experiment using Amazon's Mechanical Turk, the well-known crowd-sourcing platform. We first asked the participants to transcribe a recorded version of each sentence, as a means to give them a reference and assess their good knowledge of French. We then played the synthesized version of the same sentence, and asked the participants to rate it, in a scale from 1 (bad) to 5 (excellent) in terms of intelligibility and naturalness. We analyzed responses from 12 participants, who were able to transcribe all ten sentences correctly (with at most one phoneme wrong across the ten sentences). The mean score for intelligibility was 2.58 (standard deviation 1.21) and that for naturalness 2.84 (1.00). A breakdown of these scores across the ten sentences is shown in Table 1. These scores are not very high but considered encouraging given the novelty and experimental nature of the method.

## 5. Concluding remarks

Maeda's simulation of the vocal-tract system works under several simplifying approximations, and has been used mostly for the study of static vocal-tract configurations. Our results indicate that, (i) if driven appropriately, this simple model can be used effectively to synthesize connected speech, but (ii) there is significant room for improvement.

In the future, we plan to explore the possibility of synthesizing speech directly from contours tracked in real-time MRI data of the vocal tract [17], which would eliminate the need for a model of the geometry of the vocal tract, and offer even more realistic articulatory dynamics. The same real-time MRI data could also be used to build new geometrical models, especially for languages other than French, for which Maeda's geometrical model was originally constructed. Finally, important questions regarding the relationship between the 2D and the 3D geometry of the vocal tract could be answered using volumetric 3D MRI data [18].

## 6. Acknowledgments

We are grateful to Shinji Maeda, Yves Laprie and Slim Ouni, for mentoring, sharing code and data, and contributing to earlier versions of the methods presented in this paper. This work was supported by NIH grant DC007124.

# 7. References


[1] D. Klatt, "Review of text-to-speech conversion for English," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[2] 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.

[3] 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.

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

[5] S. Öhman, "Coarticulation in VCV utterances: spectrographic measurements." *Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[6] 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.

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

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

[9] A. Toutios, S. Ouni, and Y. Laprie, "Estimating the parameters of an articulatory model from electromagnetic articulograph data," *Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3245–3257, 2011.

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

[11] S. Maeda, "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.

[12] A. Zierdt, P. Hoole, M. Honda, T. Kaburagi, and H. Tillmann, "Extracting tongues from moving heads," in *5th Speech Production Seminar*, Kloster Seeon, Germany, 2000, pp. 313–316.

[13] A. Bothorel, P. Simon, F. Wioland, and J. Zerling, *Cinéradiographie des voyelles et consonnes du français*. L'Institut de Phonétique de Strasbourg, France, 1986, pp. 38–45,74–81.

[14] G. Fant, "Vocal source analysis-a progress report," *STL-QPSR*, vol. 20, no. 3-4, pp. 31–53, 1979.

[15] S. Maeda, "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] J. Flanagan, K. Ishizaka, and K. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell System Technical Journal*, vol. 54, no. 3, pp. 485–506, 1975.

[17] 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.

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