Protocol for a Model-based Evaluation of a Dynamic Acoustic-to-Articulatory Inversion Method using Electromagnetic Articulography

Asterios Toutios, Slim Ouni and Yves Laprie

INRIA Lorraine - CNRS UMR7503 - Nancy-Université
615, rue du Jardin Botanique, 54602 Villers-lès-Nancy, France
E-mail: {asterios.toutios, slim.ouni, yves.laprie}@loria.fr

Abstract

Acoustic-to-articulatory maps based on articulatory models have typically been evaluated in terms of acoustic accuracy, that is, the distance between mapped and observed acoustic parameters. In this paper we present a method that would allow for the evaluation of such maps in the articulatory domain. The proposed method estimates the parameters of Maeda's articulatory model on the basis of electromagnetic articulograph data, thus producing full midsagittal views of the vocal tract from the positions of a limited number of sensors attached on articulators.

1 Introduction

Our previously published acoustic-to-articulatory inversion method [7] relies on Maeda's articulatory model [5, 6], which describes the vocal tract shape in the form of a weighted sum of seven linear components. The inversion method is based on a codebook that allows obtaining the acoustic image of a vocal tract shape represented by its seven articulatory parameters. The codebook is composed of a list of paired data, each of them associating one articulatory set of parameters with its acoustic image in terms of formants or spectral contours. During inversion, the codebook enables recovering articulatory parameters that correspond to the observed acoustic input. A nonlinear smoothing algorithm together with a regularization technique is then used to recover the articulatory trajectory.

The assessment of computational acoustic-to-articulatory maps has most often focused on acoustic accuracy, which is the distance between mapped and observed formant frequencies or between map-generated and observed spectra. Since the goal is to obtain articulatory trajectories as those produced by the speaker, we propose the assessment of the quality of inversion by means of comparing the obtained articulatory trajectories with the actual dynamics of the vocal tract. Dynamics is measured using a Carstens AG500 Electromagnetic Articulograph [10] which allows the positions of several sensors to be tracked at a frequency equal to 200Hz.

The problem in exploiting EMA measurements in the context of evaluating model-based maps rises from the fact that they concern only a limited number of sensors attached on articulators, whereas the models typically describe articulation by sets of parameters which generate full (albeit simplified) sagittal vocal tract profiles. We address this problem, for the case of Maeda's model, by presenting a process that uses the EMA measurements to estimate the corresponding model parameters.

2 Speaker adaptation of the model

Maeda's model uses a semipolar grid system which consists of a polar part and two linear parts for the buccal and pharyngeal areas. It was initially built for the female speaker PB and has to be adjusted for our speaker YL.

Two adjustments are necessary. First, the determination of the mouth and pharynx length scale factors. Second, the correction of the contour of the fixed part of the vocal tract. These two steps are achieved manually by superimposing the grid system on an MRI picture of the speaker. Figure 1 shows the result of this process. The scale factors are determined to be 1.3 for the mouth and 1.2 for the pharynx.

3 Registration of EMA data

In our EMA recording setup, three reference sensors were used for head movement correction. They were placed on rigid structures, on the bridge of the nose and the two temporal bones right behind the

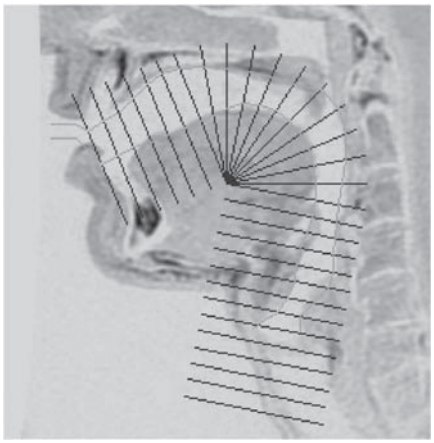

Figure 1: Adapted grid superimposed on an MRI picture of YL during the production of /a/.

ears. Four sensors were glued on the tongue, on a line roughly on the midsagittal plane, in equal distances between the apex and the area roughly under the velum. Two sensors were placed on the lip corners, one on the middle of the upper lip, one on the middle of the lower lip, and one on the lower incisors.

We acquired 3-dimensional positions of the sensors, sampled at 200 Hz, for a corpus consisting of 36 VV sequences, 100 VCV sequences and 136 continuous French middle-sized sentences. Additionally, at the end of the recording session, we used one sensor to draw a trace of the palate.

In order to register these data in a system relevant to Maeda's semipolar grid two steps are necessary. First, the definition of a midsagittal plane (in the EMA measurement coordinate system) and the subsequent projection of the data onto it. Second, the translation and rotation of the midsagittal plane coordinate system so that it matches the system defined by the grid.

The three reference sensors define a transverse plane. The midsagittal plane is defined as the plane perpendicular to the transverse one, that passes through the position of the nose bridge sensor and the midpoint between the two ear sensors.

The EMA trace of the palate is registered to the external vocal tract of the model by an automatic procedure that uses an Iterative Closest Point algorithm [9]. This results in a set of parameters for translation and rotation that are then applied to all EMA data (see Figure 2).

4 Extraction of articulatory variables

Madea's model defines a set of variables, that is, articulatory measurements performed on the grid.

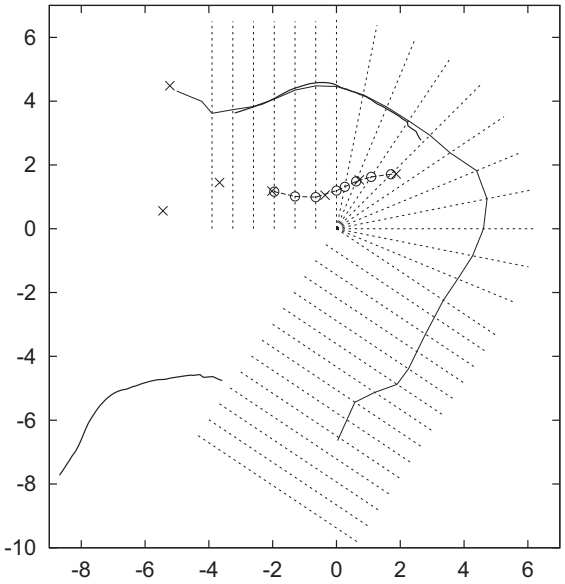

Figure 2: Original trace of palate (lower left corner); registered trace of palate; registered EMA data for tongue, lower incisor, upper and lower lip sensors (crosses); and interpolated tongue contour (circles). Axis labels denote centimeters.

Our task is to extract them on the basis of the EMA measurements.

The EMA information on the tongue consists of only the positions of four fleshpoints. We interpolate these positions by cubic splines [8] in order to derive the values of the tongue contour at the gridlines. These, after proper scaling, are the *tongue variables*. The four sensors on the tongue cover only a limited area on the grid, thus interpolation leads to the values of the contour on a limited number of gridlines, typically 6 to 8 (see Figure 2).

The definitions of the variables corresponding to the jaw and lips given by Maeda are adapted to better relate to the available EMA information. The jaw variable is defined as the projection of the position of the lower incisor sensor on its first principal component. The lip opening variable is defined as the distance between the upper and lower lip sensor positions, projected on the vertical axis of the buccal linear part of the grid. The lip protrusion variable is defined as the midpoint between the upper and lower lip sensor positions, projected on the horizontal axis of the buccal linear part of the grid. The lip width variable is defined as the distance between the two sensors on the lip corners, projected on an axis perpendicular to the midsagittal plane. The lip and jaw variables are converted to z-scores using statistics calculated from the full recording session.

5 Estimation of model parameters

In the model, the variables already described are generated from a set of underlying parameters via a linear table:

$$\mathbf{v} = \mathbf{A}\mathbf{p} \tag{1}$$

where $\mathbf{v}$ is a 29-dimensional vector (consisting of the jaw variable, the three lip variables and 25 tongue variables); $\mathbf{A}$ is a 29×6 table of real values (actually a concatenation of the "lip" and "tongue" tables in the definition of the model); and $\mathbf{p}$ is a 6-dimensional vector of parameters (jaw position, tongue body position, tongue shape, apex position, lip opening and lip protrusion). The model's larynx parameter is excluded from our present discussion, since no information on the state of the larynx may be derived from the EMA measurements.

Given a subset of measured variables, say C, at a given time instant, we can approximate the corresponding parameters by minimizing the quantity:

$$I_s = \sum_{i \in C} \left(v_i - \sum_{j=1}^{6} a_{i,j} p_j \right)^2 \tag{2}$$

where the $a_{i,j}$ are elements of matrix $\mathbf{A}$ and the p_j elements of vector $\mathbf{p}$. We further introduce the constraints:

$$p_j \in [-3, 3],\ j = 1, \ldots, 6 \tag{3}$$

which ensure that the vocal tract configurations corresponding to the estimated parameters are realistic. The minimization of Eq. (2) subject to Eq. (3) constitutes a regular constrained quadratic programming problem that is solved using an active space null set method [3].

To ensure the smoothness of the articuatory trajectories over time, we use a regularization technique based on variational calculus [2]. We introduce the following cost function to be minimized over the time interval $[t_s, t_f]$:

$$\begin{aligned} I_d = & \int_{t_s}^{t_f} \sum_{i \in C} \left(v_i(t) - \sum_{j=1}^{6} a_{i,j} p_j(t) \right)^2 dt \\ & + \lambda \int_{t_s}^{t_f} \sum_{j=1}^{6} p'_j(t)^2 dt + \beta \int_{t_s}^{t_f} \sum_{j=1}^{6} p_j(t)^2 dt \end{aligned} \tag{4}$$

The first integral of this cost function expresses the proximity between observed variables and those generated by the model parameters. The second integral expresses the changing rate of articulatory parameters. The third integral is an energy term that penalizes large articulatory efforts and prevents the vocal tract from reaching positions too far from the equilibrium. Constants λ and β are chosen heuristically.

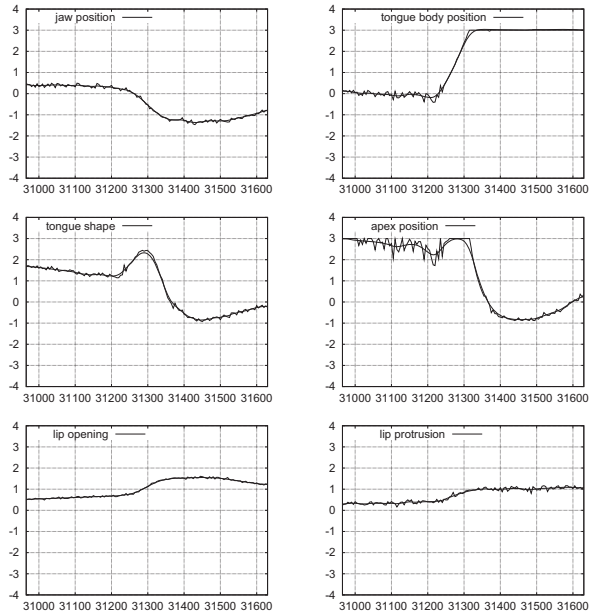

Figure 3: Articulatory trajectories for the sequence /ia/. In each graph the trajectory obtained by framewise optimization and that obtained by using the variational regularization method (smoothest trajectories) are plotted. The horizontal axis represents time in milliseconds.

The minimization of I_d gives rise to an iterative process that updates the parameter values at each step, until the satisfaction of a convergence criterion [4]. The articulatory parameter vectors found by minimizing Eq. (2) subject to Eq. (3) for the discrete-time samples in the interval $[t_s, t_f]$ are used as a startup solution for this iterative process.

6 Results and conclusion

We present a result of our procedure for the utterance of sequence /ia/ by YL. Figure 3 shows the approximated articulatory trajectories for this sequence. Figure 4 shows the evolution of the vocal tract shapes derived on the basis of these trajectories, with the larynx parameter fixed at a zero value. These shapes offer a means of validating the procedure, since one can tell whether they agree with phonetic knowledge.

Another way to validate the procedure is to synthesize speech on the basis of the derived articulatory trajectories and compare with the concurrently recorded actual speech. Figure 5 shows the three first formants derived from the articulatory sequence of

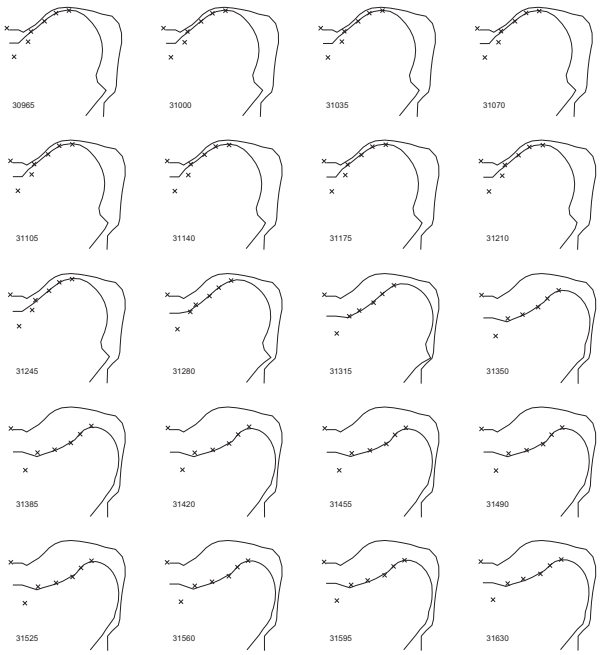

Figure 4: Temporal dynamics of the vocal tract shapes for the VV sequence /ia/. EMA sensors on the tongue, lower incisor, upper and lower lip are marked with crosses. The number at the bottom-left corner of the images represents time in milliseconds.

Figure 3 for three fixed values of the larynx parameter (-3,0 and +3) superimposed on a spectrogram of the corresponding actual speech. The matching is not perfect, which is a matter under further analysis.

The actual evaluation of our inversion method is still to be performed. However, we have developed a tool that will allow a direct comparison of articulatory trajectories derived by inversion against those corresponding to the actual vocal tract dynamics, as recorded by EMA.

Beside inversion evaluation, we may envision additional benefits from our proposed method. The ability to drive an articulatory model on the basis of EMA data may be of great potential interest for applications like speech synthesis or visualization. Future work includes applying similar ideas in cases of different articulatory models and/or different articulatory data (e.g. ultrasounds [1]).

Acknowledgment

We acknowledge the financial support of Région Lorraine for acquiring the Electromagnetic Articulograph.

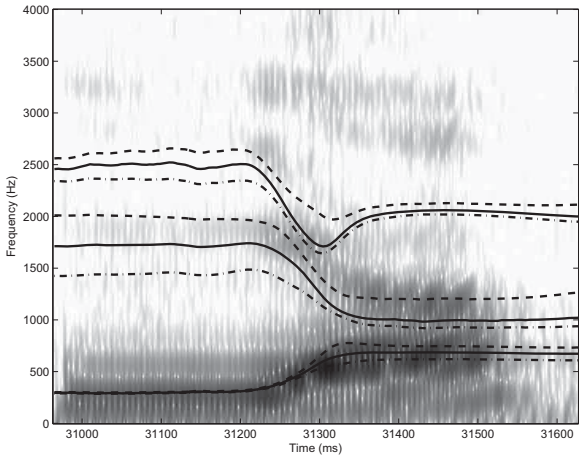

Figure 5: First three formants of /ia/ derived from the six trajectories of Fig.3 with the larynx parameter equal to zero (solid lines), -3 (dotted-dashed lines), +3 (dashed lines), superimposed on the spectrogram of actual speech.

References

[1] M. Aron, E. Kerrien, M. Berger, and Y. Laprie. Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition setup and preliminary results. In *Int. Seminar on Speech Production*, pages 435–442, 2006.

[2] M. Bonvalet. *Les principes variationnels*. Masson, 1993.

[3] R. Fletcher. *Practical methods of optimization*. Wiley-Interscience New York, NY, USA, 1987.

[4] Y. Laprie and B. Mathieu. A Variational Approach for Estimating Vocal Tract Shapes from the Speech Signal. In *Proc. ICASSP*, volume 2, pages 929–932, Seattle, Washington, USA.

[5] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10emes Journees d' Etude sur la Parole*, pages 152–162, May 1979.

[6] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 131–149. 1990.

[7] S. Ouni and Y. Laprie. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 118(1):444–460, July 2005.

[8] C. Ueberhuber. *Numerical Computation: Methods, Software, and Analysis*. Springer, 1997.

[9] Z. Zhang. Iterative point matching of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994.

[10] A. Zierdt, P. Hoole, M. Honda, T. Kaburagi, and H. Tillmann. Extracting tongues from moving heads. In *Proc. 5th Speech Production Seminar*, pages 313–316, 2000.