A VARIATIONAL APPROACH FOR ESTIMATING VOCAL TRACT SHAPES FROM THE SPEECH SIGNAL

Yves Laprie and Bruno Mathieu

LORIA
BP 239
54506 Vandœuvre-lès-Nancy, FRANCE
laprie,mathieu@loria.fr

ABSTRACT

This paper presents a novel approach to recovering articulatory trajectories from the speech signal using a variational calculus method and Maeda's articulatory model. The acoustic-to-articulatory mapping is generally assessed by a double criterion: the acoustic proximity of results to acoustic data and the smoothness of articulatory trajectories.
Most of the existing methods are unable to exploit the two criteria simultaneously or at least at the same level. On the other hand, our variational calculus approach combines the two criteria simultaneously and ensures the global acoustic and articulatory consistency without further optimization. This method gives rise to an iterative process which optimizes a startup solution given by an improved lookup algorithm. Codebooks generated with an articulatory model show nonuniform sampling of the acoustic space due to nonlinearities of the acoustic-to-articulatory mapping. We therefore designed an improved lookup algorithm building realistic articulatory trajectories which are not necessarily defined throughout the speech signal.

1. INTRODUCTION

Estimating the vocal tract shape from the speech signal has received considerable attention because it offers new perspectives for speech coding as well as speech recognition [8]. Most of the works on acoustic-to-articulatory inversion rely on articulatory models which ensure a realistic approximation of the vocal tract with a limited number of parameters.

Mapping between the acoustic and articulatory domains is non-unique. Estimating articulatory parameters therefore requires the use of codebook lookup or neural network techniques [10]. Whatever the method is, it must lead to slowly changing parameters which generate spectra as close as possible to those of the original speech. This corresponds to satisfying two criteria: proximity to acoustic data and smoothness of articulatory trajectories, which are therefore required to evaluate the quality of the acoustic-to-articulatory mapping. Generally, existing methods cannot allow for the two criteria at the same level, or at least favor one criterion to the detriment of the other. In the case of the mapping via articulatory codebooks, for example, the acoustic distance is used to retrieve vocal tract shapes. Then, dynamic constraints are imposed on the evolution of articulatory parameters. Finally, a local optimization improves the acoustic proximity with the input signal.

In this paper we propose a new method of combining the two criteria. This method utilizes the well known theory of variational calculus [9] which gives rise to an iterative process. This process starts with an initial solution and generates a sequence of articulatory trajectories which optimizes a cost function which combines acoustic distance and changing rate of articulatory parameters.

There are three major advantages of this method compared to many other existing methods:

- it involves the continuous nature of articulatory trajectories and the global acoustic and articulatory consistency without further optimization.
- it incorporates the acoustic behavior of the articulatory model by means of sensitivity functions of formants, with respect to articulatory parameters.
- it is possible to investigate how rough inversion solutions can be modified to minimize the cost function. This allows compensatory effects to be studied, by considering startup articulatory trajectories corresponding to different relative positions of articulators.

2. FORMANT TO ARTICULATORY MAPPING WITH VARIATIONAL CALCULUS METHOD

Firstly, we present the variational method applied to acoustic-to-articulatory inversion. Maeda's articulatory model [4] describes the vocal tract shape by means of seven parameters. These parameters are time functions $\alpha(t) = (\alpha_1(t) \dots \alpha_i(t) \dots \alpha_7(t))$, $t \in [t_i, t_f]$. Formant trajectories extracted from speech $f_j(t)$, $1 \leq j \leq 3$ are the input data. Those generated by the acoustic simulation are $F_j(\alpha(t))$ ($1 \leq j \leq 3$). A cost function for evaluating acoustic-to-articulatory mapping incorporates two components:

- $\sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t)))^2$ which expresses the proximity between observed acoustic data, i.e. formants trajectories $f_j(t)$, and those generated by the model $F_j(\alpha(t))$.
- $\sum_{i=1}^{7} m_i \alpha_i'^2(t)$ which expresses the changing rate of articulatory parameters. In order to penalize large articulatory efforts and prevent the vocal tract from reaching positions too far from equilibrium, a potential energy term $\sum_{i=1}^{7} k_i \alpha_i^2(t)$ is added.

The cost function to be minimized has the following form

$$I = \int_{t_i}^{t_f} \sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t)))^2 dt + \lambda \int_{t_i}^{t_f} \sum_{i=1}^{7} m_i \alpha_i'^2(t) dt + \beta \int_{t_i}^{t_f} \sum_{i=1}^{7} k_i \alpha_i^2(t) dt \quad (1)$$

where t_i and t_f define the time interval over which the inversion is carried out, λ and β express the compromise between the changing rate of articulatory parameters, their distance from equilibrium and the acoustic distance. m_i is the pseudo mass of the ith articulator, and k_i is its spring constant. Eq. 1 can be written as

$$I = \int_{t_i}^{t_f} \Phi(\alpha(t), \alpha'(t), t)dt$$

Variational calculus [9] can be used to minimize I. Euler-Lagrange equations express the vanishing of the derivative of I with respect to each of the α_i. These equations are a necessary condition to insure a minimum of I and can be written

$$\begin{cases} \frac{\partial \Phi}{\partial \alpha_1} - \frac{d}{dt}\frac{\partial \Phi}{\partial \alpha_1'} = 0 \\ \dots \\ \frac{\partial \Phi}{\partial \alpha_7} - \frac{d}{dt}\frac{\partial \Phi}{\partial \alpha_7'} = 0 \end{cases} \quad (2)$$

Considering the definition of Φ, each of the Euler-Lagrange equations becomes:

$$\sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t)))\frac{\partial F_j}{\partial \alpha_i} + \beta k_i \alpha_i(t) - \lambda m_i \alpha_i''(t) = 0 \quad i = 1 \dots 7 \quad (3)$$

where $\alpha_i''(t)$ is the second time derivative of $\alpha_i(t)$. From now on we only consider one of the equations of the system (2) for sake of clarity. We assume that we have a rough startup estimation of the articulatory trajectory α_i (see section 3). We can therefore define an iterative process $\alpha_i^\tau(t)$ such that

$$\lim_{\tau \to \infty} \alpha_i^\tau(t) = \alpha_i(t)$$

(where $\alpha_i^{\tau=0}(t)$ is the startup solution) using the associated evolution equation

$$\gamma \frac{\partial \alpha_i^\tau}{\partial \tau} + \beta k_i \alpha_i^\tau - \lambda m_i \alpha_i^{\tau\prime\prime} = -\sum_{j=1}^{3}(f_j(t) - F_j(\alpha^\tau(t)))\frac{\partial F_j}{\partial \alpha_i^\tau} \quad (4)$$

$\frac{\partial \alpha_i^\tau}{\partial \tau}$ represents the evolution of parameter α_i during the iteration process and γ a parameter for controlling the evolution rate. A solution to the static equation Eq. 3 is found when the term $\gamma \frac{\partial \alpha_i^\tau}{\partial \tau}$ vanishes.

For sake of convenience we set m and k to 1. This choice will be questioned later in the paper.

Let $\boldsymbol{\alpha^\tau} = (\alpha_{i,0}^\tau, \dots \alpha_{i,k}^\tau, \dots \alpha_{i,N}^\tau)$ denote the discrete representation of $\alpha_i(t)$, $\alpha_{i,k}^\tau$ represents the value of α_i^τ at discrete time $t = t_i + k\frac{t_f - t_i}{N}$ in the iteration τ. Since solving Eq. 4 for α_i is independent of other articulatory trajectories, $\alpha_{i,k}^\tau$ is noted α_k^τ for sake of clarity.

Let $(f_0, \dots f_k, \dots f_N)$ denote the observed formant trajectory and $(F_0, \dots F_k, \dots F_N)$ the formant trajectory generated by the acoustic simulation. Finite difference approximation of the derivative $\alpha''(t)$ leads to the following equation

$$\begin{aligned} \gamma(\alpha_k^\tau - \alpha_k^{\tau-1}) + \beta\alpha_k^\tau - \lambda(\alpha_{k+1}^\tau - 2\alpha_k^\tau + \alpha_{k-1}^\tau) \\ = -\sum_{j=1}^{3}(f_{j,k} - F_{j,k})\left.\frac{\partial F_j}{\partial \alpha}\right|_{\alpha_{1,k}^\tau \dots \alpha_{7,k}^\tau} \end{aligned} \quad (5)$$

where τ represents the iteration under process and k the discrete time. The derivative term $\left.\frac{\partial F_j}{\partial \alpha}\right|_{\alpha_{1,k}^\tau \dots \alpha_{7,k}^\tau}$ is calculated for the parameter α_i at point $(\alpha_{1,k}^\tau \dots \alpha_{7,k}^\tau)$ and incorporates the behavior of the acoustic modeling with respect to the evolution of articulatory parameters.

Boundary conditions are needed to ensure that Eq. 5 has a unique solution. Since we do not impose any constraint on the positions of the extremities of $\alpha(t)$

$$\alpha'(0) = \alpha'(N) = 0$$

are the boundary conditions. Let B be an $(N+1) \times (N+1)$ matrix

$$B = \begin{bmatrix} \gamma+\beta+\lambda & -\lambda & 0 & \cdots & 0 \\ -\lambda & \gamma+\beta+2\lambda & -\lambda & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\lambda & \gamma+\beta+2\lambda & -\lambda \\ 0 & \cdots & 0 & -\lambda & \gamma+\beta+\lambda \end{bmatrix}$$

$$\boldsymbol{\alpha^\tau} = (\alpha_0^\tau, \dots \alpha_k^\tau, \dots \alpha_N^\tau)^T$$

$$\boldsymbol{c^\tau} = \begin{bmatrix} \gamma\alpha_0^{\tau-1} - \sum_{j=1}^{3}(f_{j,0} - F_{j,0})\frac{\partial F_j}{\partial \alpha} \\ \gamma\alpha_1^{\tau-1} - \sum_{j=1}^{3}(f_{j,1} - F_{j,1})\frac{\partial F_j}{\partial \alpha} \\ \cdots \\ \gamma\alpha_N^{\tau-1} - \sum_{j=1}^{3}(f_{j,N} - F_{j,N})\frac{\partial F_j}{\partial \alpha} \end{bmatrix}$$

Eq. 5 can be put in matrix form

$$B\boldsymbol{\alpha^\tau} = \boldsymbol{c^\tau}$$

Solving Eq. 2 leads to an iterative process. At each iteration $\boldsymbol{\alpha^\tau}$ is calculated for each of the seven articulatory parameters α_i. In order to ensure that a minimal solution of Eq. 1 is reached, one needs to choose a good startup solution. This is achieved by employing a method derived from codebook lookup. The startup solution is then iteratively transformed so that Eq. 1 is minimized.

3. CHOICE OF A STARTUP SOLUTION

As the cost function defined in Eq. 1 is non-convex we need a startup solution set somewhere near a relevant minimum of I. Preliminary experiments have shown that it is possible to approximate formant trajectories very accurately, as soon as the codebook gets very large. However, this generally does not ensure that articulatory trajectories are smooth and relevant, at least when the constraints imposed are not too stringent. Several reasons may explain why tract shapes at t and $t-1$ may be very far away from each other:

- the codebook sampling is insufficiently fine. There is therefore no articulatory entry corresponding to the formants observed which is close to the entry retrieved at $t-1$. This may happen near a mapping nonlinearity, i.e. when a small variation in articulatory parameters gives a large acoustic variation.
- the articulatory model is unable to generate a tract shape identical to that produced by the subject. Formants are therefore reached by means of an unjustified compensatory effect, which gives rise to outliers in articulatory trajectories.

The second reason is all the more important since a deep adaptation of the model to a new speaker is not possible.

We prefer therefore a partial startup solution (not necessarily defined throughout the time interval where inversion is performed) rather than one which is defined over the whole speech segment,

but which is not realistic and thus difficult to optimize. The variational approach has the advantage that a partial solution, filled in by linear interpolation, is transformed and optimized by taking into account the acoustic behavior of the articulatory model.

We designed an improved codebook-lookup algorithm to construct good startup solutions. This algorithm derives from a nonlinear smoothing algorithm proposed by Ney [6]. The aim is to smooth a curve given by a set of points with a certain number of outliers. Ney proposed to use dynamic programming to choose the points which ought to be kept. In the case of acoustic-to-articulatory mapping this leads to the choosing of a sequence of shapes, possibly not defined at each instant of the inversion.

Firstly, a set of tract shapes $s(i)$ is retrieved at each instant, each of these shapes produces formants close to those of the original spectrum. The purpose of the lookup algorithm is to find an articulatory trajectory formed by these shapes in the sequence of the shape sets:

$$S = (s(0) \dots s(i) \dots s(N))$$

where i is the discrete time. The construction of a trajectory gives rise to a double selection:

- the choice of instants at which the trajectory is defined, i.e. the choice of a subsequence of S defined by a function j:

  $$\overline{S} = (s(j(0)) \dots s(j(k)) \dots s(j(K)))$$

  where $K < N$ and j is a monotonic function: $0 \leq j(k) < j(k+1) \leq N$.
- the choice of one shape in each of the sets selected $s(j(0)) \dots s(j(k)) \dots s(j(K))$. The shape chosen out of the set $s(j(k))$ is denoted $\alpha(j(k))$ ($\alpha(j(k)) \in I\!R^7$) and the articulatory trajectory is therefore

  $$\overline{A} = (\alpha(j(0)) \dots \alpha(j(k)) \dots \alpha(j(K)))$$

The cost of choosing $\alpha(j(k))$ after $\alpha(j(k-1))$ derives directly from Eq.1

$$C(\alpha(j(k)), \alpha(j(k-1))) = \sum_{j=1}^{3} (f_j(t) - F_j(\alpha(j(k))))^2 + \lambda \sum_{i=1}^{7} m_i (\alpha_i(j(k)) - \alpha_i(j(k-1)))^2 + \beta \sum_{i=1}^{7} k_i \alpha_i(j(k))^2$$

Based on this local cost, an overall cost function to be minimized could be

$$\sum_{j=1}^{K} C(\alpha(j(k)), \alpha(j(k-1)))$$

However, since the local cost is positive, a solution which minimizes this function could be reduced to a very small number of shapes, possibly zero, which is unacceptable. In the case of nonlinear smoothing, Ney proposed to introduce a nonnegative bonus B which represents the interest of preserving a shape in the final trajectory. The overall cost function becomes

$$D = \sum_{j=1}^{K} (C(\alpha(j(k)), \alpha(j(k-1))) - B)$$

As proposed by Ney this problem is efficiently solved by dynamic programming.

The minimization of the cost function D by dynamic programming is obtained through the calculation of partial measures of D for each shape in each set of shapes $s(i)$. This calculation involves the examination of all the shapes in all the sets of shapes which come before $s(i)$. Clearly, a vast number of cases examined are unnecessary because a reasonable solution should cover most of the speech segment under investigation without large gaps. We therefore, only explore sets of shapes within a moving window before $s(i)$, which significantly reduces the computational time.

4. EXPERIMENTS

We are using the linear component articulatory model of Maeda. A series of MRI images was realized to adapt the model to our subject [5]. We modified scale factors of pharynx and mouth cavities and determined the wall (the motionless contour of the vocal tract). Linear components are kept unchanged because they result from a statistical analysis of a vast X-ray database. Despite the adaptation, one cannot ensure that the model will perfectly approximate area functions that our subject will generate.

As proposed in [3] codebooks can be constructed by either the root shape interpolation method or the random sampling method. A "root shape" corresponds to the articulatory configuration of a given steady vowel. Entries of the codebook are obtained by sampling along a trajectory from one root shape to another. Although the scattering of acoustic-to-articulatory mapping is substantially smaller (see [5]) with the "root shape" method, we accepted the random sampling method because some vowels were strongly centralized by our subject (probably because of the noise produced by the MRI machine). A codebook of 300,000 entries was generated and the entries of the codebook not corresponding to realistic vocalic area functions were filtered out using constraints proposed by Boë et al.[1].

Fig. 1 and Fig. 2 show the formant trajectories generated by the articulatory parameters recovered from the formant trajectories, which were extracted from speech, and three articulatory parameter trajectories (jaw position, tongue dorsum position and lip aperture for Fig. 1, jaw position, tongue dorsum position and tongue shape for Fig. 2). Formant trajectories were extracted by means of our formant tracking algorithm [2]. In fact, we could easily use spectral vectors as acoustic data instead of formant trajectories. In this case, formants would be used as a key to access the codebook. In the two examples, formant trajectories superimpose well on spectral peaks and the articulatory trajectories are relevant with the sounds uttered.

In some cases, especially when /a/ follows a closed vowel (as in Fig. 2 for instance) an exaggerated compensatory effect between jaw position and lip aperture is observed. It is due to the minimization of the term $\sum_{i=1}^{7} m_i \alpha_i'^2(t)$ which can favor the minimization of the lip parameter rather than that of the jaw, according to the initial articulatory parameters. This stems from the fact that pseudo masses and spring constants have been set to 1 which is unrealistic since the inertia of the jaw is much larger than that of lips. Pseudo mass and spring constant values need to be learned or evaluated on the basis of physiological measures to prevent unjustified compensatory effects. Another solution to penalize co-occurrences of a low lip aperture and a very open jaw position would consist of incorporating a term $\alpha_{lip}\alpha_{jaw}$ in Eq. 1.

In spite of necessary improvements dealing with parameters m_i and k_i, our method exhibits qualities of flexibility which make possible the investigation of real compensatory effects. Our al-

gorithm, for determining startup solutions, turns out to be very efficient since it eliminates outliers which usually lead to chaotic articulatory parameters.

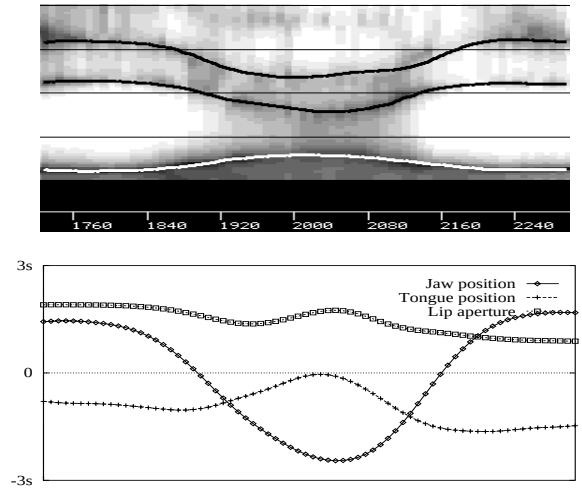

Figure 1: Formant trajectories generated by articulatory trajectories recovered from /iai/ and three articulatory parameters representing deviations from the neutral position.

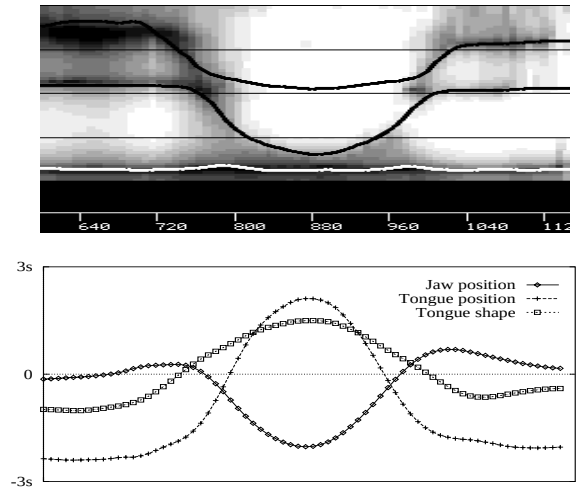

Figure 2: Formant trajectories generated by articulatory trajectories recovered from /iui/ and the three articulatory parameters representing deviations from the neutral position.

5. CONCLUDING REMARKS

Unlike most other techniques for recovering articulatory trajectories from acoustic data, our method exhibits a dynamic behavior: the startup solution deforms itself according to the sensitivity functions of formants with respect to articulatory parameters.

This approach has the advantage that it does not impose constraints on startup solutions, which are too stringent. Consequently, it is well suited for investigating compensatory and coarticulation effects, especially when EMG data are available to compare real and recovered articulatory trajectories.

The current system incorporates the acoustic behavior of Maeda's articulatory model. The dynamic behavior of the vocal tract is partly governed by pseudo masses m_i and spring constants k_i which have been arbitrarily set to 1. We currently investigate how cross-validation methods [7] could be exploited to optimize these parameters.

Acknowledgments

We would like to thank Dr. Shinji Maeda for making his articulatory model available and Dr. Marie-Odile Berger for fruitful discussions.

6. REFERENCES

[1] L.-J. Boë, P. Perrier, and G. Bailly. The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20:27–38, 1992.

[2] Y. Laprie and M.-O. Berger. Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication*, 19(4):255–270, October 1996.

[3] J. N. Larar, J. Schroeter, and M. M. Sondhi. Vector quantization of the articulatory space. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-36(12):1812–1818, December 1988.

[4] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

[5] B. Mathieu and Y. Laprie. Adaptation of Maeda's model for acoustic to articulatory inversion. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 4, pages 2015–2018, Rhodes, Greece, September, 1997.

[6] H. Ney. A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, 5(2):163–173, March 1983.

[7] H. B. Richards, J. S. Bridle, M. J. Hunt, and J. S. Mason. Dynamic constraint weighting in the context of articulatory parameter estimation. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 5, pages 2535–2538, Rhodes, Greece, September, 1997.

[8] R.C. Rose, J. Schroeter, and M.M. Sondhi. An investigation of the potential role of speech production models in automatic speech recognition. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 575–578, Yokohama, Japan, September, 1994.

[9] R.S. Schechter. *The variational Method in Engineering*. McGraw-Hill Book Comp., New York, 1967.

[10] J. Schroeter and M. M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. on Speech and Audio Processing*, 2(1, Part. II):133–150, January 1994.