A robust variational method for the acoustic-to-articulatory problem

Blaise Potard, Yves Laprie

INTERSPEECH
2009 BRIGHTON

A robust variational method for the acoustic-to-articulatory problem

Blaise Potard.[1], *Yves Laprie*[1]

[1]LORIA, CNRS / Nancy Université, Nancy, France

potard@loria.fr, laprie@loria.fr

Abstract

This paper presents a novel acoustic-to-articulatory inversion method based on an articulatory synthesizer and variational calculus, without the need for an initial trajectory. Validation in ideal conditions is performed to show the potential of the method, and the performances are compared to codebook based methods. We also investigate the precision of the articulatory trajectories found for various acoustic vectors dimensions. Possible extensions are discussed.

Index Terms: acoustic-to-articulatory inversion, variational calculus.

1. Introduction

Acoustic-to-articulatory inversion, i.e. the recovery of the vocal tract configuration from the speech signal, is a long standing problem in speech research. Over the last 40 years, many different approaches have been proposed to address it, that can roughly be classified into model-based and data-based approaches. In the recent years, most studies have been focusing on data-based methods (such as the ones presented in [1, 2]), which tend to give more accurate results.

Although model-based methods were much more widely investigated in the past [3], research is still active in that domain. Among model-based approaches, several inversion methods [4, 5] use variational calculus as a post-processing step to smooth an initial trajectory, but to our knowledge, it has never been used as a stand-alone acoustic-to-articulatory inversion method.

The method presented here is based on analysis by synthesis: an articulatory model controlled by control parameters allows us to mimic the deformations mode of a real human speaker, and an articulatory synthesizer allows us to generate the corresponding sound by computing the area function. The coupling of the articulatory model and the synthesizer provide a “black box” F, from which for any articulatory vector α representing a set of command for the model we can compute an acoustic vector $F(\alpha)$.

In our case, the articulatory model / synthesizer used are Maeda’s articulatory model [6] and Maeda’s synthesizer[7]. An articulatory vector is a 7-dimensional real vector, each component begin typically in the interval [-3;3]. In this study, the acoustic vectors used will be the first resonances frequencies of the vocal tract configurations.

2. General presentation and motivation

In the context of the ASPI project [8], a novel method of inversion based on variational calculus was developed to solve the dynamic acoustic-to-articulatory mapping without the need for an initial trajectory.

It finds its roots in a work previously done by Laprie and Mathieu [9], but extends it so it does not require any initial trajectory to find a solution. This allows us to skip the first two steps of inversion: generation of solutions for the static acoustic-to-articulatory mapping problem using a codebook, and generation of an initial trajectory among this set of solutions using dynamic programming or non-linear filtering.

These first two steps are indeed the most time consuming, although some progress has recently been made [10] to improve the speed (and in a lesser extent, the accuracy) of the codebook inversion procedure. This procedure also requires the construction of a different codebook for each speaker, which takes a very long time. The second step has a complexity which depends on the accuracy we wish to achieve for the initial trajectory, but it is limited in any case by two main elements: the acoustic accuracy of the codebook, and the density of solutions generated during the codebook search procedure.

The third step is a classical minimization of a cost function using a variational approach, which reduces both kinds of inaccuracies from the initial trajectory. The cost function that Laprie and Mathieu proposed had the following form:

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

in which t_i and t_f are respectively the time of beginning and end of the sequence, $\alpha(t)$ is the articulatory vector of the trajectory at time t, $f(t)$ is the input acoustic vector that we wish to inverse, and $F(\alpha(t))$ is the acoustic image of the articulatory vector. Each acoustic vector has M components, which in our case corresponds to the M first resonances of the vocal tract. λ and β are weight coefficients for the two articulatory terms. This cost function is thus the combination of three basic terms:

- $\sum_{j=1}^{M}(f_j(t) - F_j(\alpha(t))^2$ expresses the proximity between observed acoustic vectors – in our case, the M first formants frequencies – and those generated by the model – in our case, the M first resonances of the transfer function. This helps reduce the acoustic imprecision of the codebook search.
- $\sum_{i=1}^{7} m_i \alpha_i'^2(t)$ expresses the changing rate of articulatory parameters; this corresponds to a kinetic pseudo-energy term, and helps reduce the sampling effect of the codebook search.
- $\sum_{i=1}^{7} k_i \alpha_i^2(t)$ expresses the distance from the neutral positions of the articulators; this term prevents the vocal tract from reaching positions too far from the equilibrium.

The variational calculus minimizes this cost function by transforming it into the following iteration:

$$\begin{aligned}-\sum_{j=1}^{M}(f_j(t) - F_j(\alpha^\tau(t))\frac{\partial F_j}{\partial \alpha_i^\tau}(t) = \\ -\lambda m_i {\alpha_i^\tau}''(t) + \beta k_i \alpha_i^\tau(t) + \gamma \frac{\partial \alpha_i^\tau}{\partial \tau}(t) \quad (2)\end{aligned}$$

Variational calculus theorems guarantee the convergence towards a minimum of the cost function, provided the initial trajectory is "close enough" from this minimum. The initial motivation for this work was to quantify this "close enough" term, to investigate whether it would be possible to skip the determination of the initial trajectory.

3. Initial study and proposed modifications

By observing the behaviour of the variational method with initial articulatory trajectories increasingly far from the original, our initial finding was that the "close enough" was usually not very far, due to the form of the acoustic term: an initial error of 20 Hz was often enough to lead to a diverging trajectory. On the other hand, both articulatory terms lead unconditionally to a stable solution; further investigation were thus conducted to modify the acoustic term and make it more stable.

In the work of Mathieu, the acoustic criterion in the cost function has the following form: $\sum_j (f_j(t) - F_j(\alpha(t)))^2$. This criterion is unfortunately prone to numerical instability, as we will explain later.

Sorokin [4] proposes to use a slightly different criterion: $\max_j |1 - F_j(\alpha(t))/f_j(t)|$ which is not very practical because this function has singular derivatives. A compromise form is the following: $\sum_j (1 - F_j(\alpha(t))/f_j(t))^2$ (which is the same as Sorokin, replacing the infinite norm by the Euclidean norm). This form is much more numerically stable than the form proposed by Laprie and Mathieu, but it is still only valid in a small vicinity of an exact solution, so it can only be used in conjunction with an initial solution. An acoustic criterion which always converges towards a correct solution would be preferable.

Let us now observe this acoustic term in the iterative form of the cost function. In the iterative form, we have the following acoustic criterion: $\sum_{j=1}^{M}(f_j(t) - F_j(\alpha^\tau(t))\frac{\partial F_j}{\partial \alpha_i^\tau}(t)$. If we assume that the current solution is not far from the expected one α^*, then we have the following: $\alpha = \alpha^* + d\alpha$, where $d\alpha$ is "small". If we assume local linearity of the articulatory-to-acoustic mapping, then $F(\alpha) = F(\alpha^* + d\alpha) = f + J \times d\alpha$, where J is the local Jacobian matrix, i.e. $\left[\frac{\partial F_j}{\partial \alpha_i}\right]_{i,j}$. To reach this optimal solution for the acoustic criterion, we would thus need to change the current vector by $d\alpha$. If we replace it within the iterative form of the acoustic criterion, we obtain:

$$\begin{aligned}-\sum_{j=1}^{M}(f_j(t) - F_j(\alpha^\tau(t))\frac{\partial F_j}{\partial \alpha_i^\tau}(t) &= J \times d\alpha^{\mathrm{T}} J \\ &= d\alpha^{\mathrm{T}} J^{\mathrm{T}} J. \\ &\approx ||J||^2 \times d\alpha\end{aligned}$$

For determining an optimal correction of the articulatory vector assuming local linearity, we can see that this criterion is very inefficient, since although it is proportional to $d\alpha$, the proportionality coefficient depends on the Jacobian matrix, and therefore on the behavior of the local mapping. To guarantee an unconditional convergence, we would need to do very small steps, which would make the method extremely slow.

Indeed, it seems more profitable to use this term: $J^{-1}(f(t) - F(\alpha))$, i.e. using the pseudo-inverse of the Jacobian matrix. Assuming local linearity, this leads directly to $d\alpha$, which is the correction vector expected.

We are thus proposing to use this term as the acoustic criterion in the iterative form. Unfortunately, this new term cannot be expressed easily into the cost function.

An explicit correction vector is computed, based on the use of the pseudo-inverse computed similarly to Schoentgen [11] through Singular Value Decomposition, which allows us to explicitly compute a correction vector that is more likely to converge towards a local minimum of the acoustic error.

Although this solution is elegant, it is not robust in case of erroneous acoustic vectors, e.g. acoustic vectors that cannot be produced by the synthesizer, and the local linearity hypothesis does not hold if the correction vector $d\alpha$ found is too large anyway. Some additional constraints on the norm of the correction vector were thus introduced to limit the effects of wrong acoustic vectors: the norm of the actual acoustic correction displacement in one iteration is bounded, and the weight of the acoustic criterion is dynamically lowered to reduce the effect of large vectors.

The actual acoustic correction vector we use is the following :

$$\frac{d\alpha}{1 + |d\alpha|/M + (|d\alpha|^2/M},$$

where M is the maximum value that the correction vector should have over one iteration.

With other technical improvement (time-varying γ, λ and β coefficients...), we have an inversion method based on variational calculus that appears to converge in most cases, even with an initial solution very far from the optimal trajectory. In practice, as the initial articulatory trajectory we use a sequence of neutral shapes.

4. Experiments

To evaluate this new method, we performed inversion on synthetic acoustic sequences generated from articulatory data from the PB corpus from IPS [12], and we compared the performance of this new method to our previous method using a codebook [13], for different sizes of the acoustic vectors.

The corpus is composed of ten French sentences, uttered by a native female speaker. The data available are sequences of articulatory parameters corresponding to the midsagittal plane of the vocal tract, obtained from lateral X-ray images, at a rate of 50 images/s, and corresponding audio recording. The acoustic signal is unfortunately very noisy.

In order to properly investigate the performances of the method itself and avoid to cope with errors due to signal processing or acoustic model mismatch, we chose to perform inversion on a synthetic acoustic signal generated from the articulatory parameters sequences. This allows us to measure the potential of the method, in ideal conditions; the results are thus not representative of what we would obtain in real conditions, i.e. with acoustic parameters obtained from a natural signal.

The performances of the inversion method is measured by comparing the inverse articulatory trajectory found to the original one. The comparison is done through two metrics:

- An articulatory distance, which is simply an Euclidean

Table 1: *Articulatory distance of the inverse articulatory trajectory to the original, for the first sentence of the corpus PB. We compare the method using a codebook to this new method for different dimensions of the acoustic vector.*

Method / NF	3	4	5
Codebook inversion	0.51	0.38	0.30
Variational calculus	0.55	0.50	0.33

distance over articulatory vectors, i.e. :

$$d_1(X,Y) = \sum_{i=1} 7(X_i - Y_i)^2,$$

where X and Y are two articulatory vectors.

- A geometric distance, measured by projecting a vocal tract shapes on Maeda's grid, and by computing the square difference with the projection of the reference VT (vocal tract) shape. In other words, if we denote as $P_{X,j}$ the coordinates of the projections of the reference VT-shape on Maeda's grid, and $P_{Y,j}$ the coordinates of the projections of a given VT-shape on the grid, we compute a geometric distance using this formula:

$$d_2(X,Y) = \sqrt{\frac{\sum_{j=1}^{N} |P_{X,j} - P_{Y,j}|^2}{N}},$$

in which N is the total number of points on Maeda's grid.

Both distances have advantages and drawbacks. The articulatory distance can more easily be interpreted in a phonetic sense, since the articulatory model has been designed so that control parameters have a phonetically relevant interpretation. This design has its drawbacks: although the parameters were chosen to be orthogonal, natural compensatory effects is such that very distant articulatory vectors can lead to very close vocal tract shape geometries, and thus very close acoustics.

The geometric distance is more strongly related to the acoustics and is therefore more relevant as a measure of the error, but cannot be easily interpreted phonetically. Furthermore, this distance is not even always relevant with regards to the acoustics, since in the presence of a narrow constriction, a very small change in the geometry can lead to a large difference in the acoustics, and in the case of a wide vocal tract, some large geometric changes can have little to no acoustic effect.

5. Results and discussion

Table 1 summarizes the results obtained when doing the inversion with various acoustic vector sizes (from 3 to 5 formant frequencies) on the first sentence of the PB corpus. The metric used is d_1, i.e. the articulatory distance, and is expressed in articulatory parameters units.

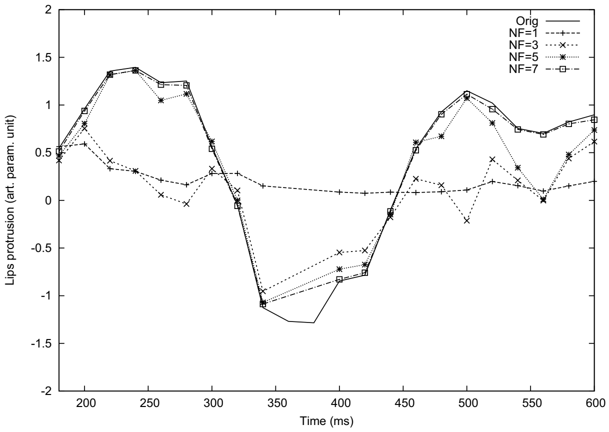
Table 2 summarizes the results obtained when doing the inversion with various acoustic vector sizes (from 3 to 5 formant frequencies) on the whole PB corpus. Both d_1 and d_2 distances are presented.

We can observe in Table 1 that the performances on the first sentence of the PB corpus are very similar with both methods, and that the distance to the original of the inverse trajectory decrease with the size of the acoustic vector. Figure 1 shows a detailed inverse trajectory with various acoustic vector sizes (from 1 to 7).

Table 2: *Average geometric (in mm) and articulatory distances (in articulatory vector units) to the original trajectory for inverse trajectory found on the whole PB corpus. The inverted signal is a synthetic signal generated from the original trajectory to avoid acoustic synthesizer mismatch.*

distance/NF	3	4	5
Articulatory	0.633	0.555	0.436
Geometric (mm)	1.77	1.42	0.92

Figure 1: Evolution of the precision of the solution found with regards to the size of the acoustic vector.



This method has proven its effectiveness for acoustic-to-articulatory inversion using the first formant frequencies as input: the accuracy is about the same as the previous method for an acoustic vector composed of the 3 first formant frequencies, which is the maximum we can expect to extract reliably from an actual acoustic signal.

The method is fairly simple, and its complexity is globally linear, but it is however quite slow due to the use of a computing-intensive articulatory synthesizer. It takes many iterations to converge towards a minimum, and each iteration requires many invocations of the articulatory synthesizer: for each sample, we compute the acoustic image of the current articulatory vector, as well as the local gradient, which requires 15 calls. For each iteration, for a sampling rate of 20ms and a sequence of 1s, we thus need 750 calls to the synthesizer. On average, about 30 iterations are required to converge, we thus need about $30 * 750 = 22500$ calls to the synthesizer. On a recent computer, it takes typically 1ms to compute an acoustic image using our synthesizer; it is therefore much slower than real time to use this method in our case.

It can however become real-time when replacing the articulatory synthesizer by a high quality codebook synthesizer such as the one presented in [14]. Unfortunately, using a codebook makes the method much less flexible, since one has to be built for each speaker; additionally, the construction time of a codebook is prohibitive.

The method still lacks robustness in the case of "mistakes" in the input acoustic vector: "impossible" acoustic vectors can sometimes produce wide errors in the articulatory trajectory, and should preferably be eliminated; this problem is even more frequent when using more complex input acoustic vectors, such as LPC coefficients.

6. Conclusions and future work

We have presented in this paper a flexible method to perform acoustic-to-articulatory inversion. This method has a linear complexity and is stand-alone, and offers performances comparable to codebook inversion methods.

This method is still experimental, and further developments need to be made. The robustness of the method to buggy acoustic vectors needs to be further addressed; additionally, the weights of the β and λ coefficients, as well as the number of iterations, have been arbitrarily fixed, but some further testing need to be done to find more appropriate –and hopefully more successful– values.

Validations on a larger corpus and with "real" acoustic data (and not synthetic signals) are ongoing and will be presented very soon.

7. References

[1] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Trans. on Speech, and Audio Processing*, vol. 12(2), pp. 175–185, 2004.

[2] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proc. INTERSPEECH*, Pittsburgh, USA, september 2006.

[3] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[4] V. Sorokin, A. Leonov, and A. Trushkin, "Estimation of stability and accuracy of inverse problem solution for the vocal tract," *Speech Communication*, vol. 30, pp. 55–74, 2000.

[5] 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*, vol. 118, no. 1, pp. 444–460, 2005.

[6] 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.

[7] ——, "Conversion of midsagittal dimensions to vocal tact area function," *Journal of the Acoustical Society of America*, 1972.

[8] S. Maeda, M.-O. Berger, O. Engwall, Y. Laprie, P. Maragos, B. Potard, and J. Schoentgen, "Acoustic-to-articulatoy inversion: Methods and acquisition of articulatory data," ASPI Consortium, Tech. Rep., November 2006.

[9] Y. Laprie and B. Mathieu, "A variational approach for estimating vocal tract shapes from the speech signal," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Seattle, USA, May 1998, pp. 929–932.

[10] B. Potard and Y. Laprie, "Improving the sampling of the null space of the acoustic-to-articulatory mapping," in *ISSP '08*, Strasbourg, France, Dec. 2008.

[11] J. Schoentgen and S. Ciocea, "Kinematic formant-to-area mapping," *Speech Communication*, vol. 21, pp. 227–244, 1997.

[12] A. Bothorel, P. Simon, F. Wioland, and J.-P. Zerling, *Cinéradiographies des voyelles et consonnes du Français*. Travaux de l'institut de Phonétique de Strasbourg, 1986.

[13] B. Potard, "Inversion acoustique-articulatoire dynamique par codebook hypercuboïque : premiers rsultats," in *JEP '08*, Avignon, France, 2008.

[14] B. Potard and Y. Laprie, "Compact representations of the articulatory-to-acoustic mapping," in *Interspeech, Anvers*, Aug. 2007.