INTRODUCTION OF CONSTRAINTS IN AN ACOUSTIC-TO-ARTICULATORY INVERSION METHOD BASED ON A HYPERCUBIC ARTICULATORY TABLE

Yves Laprie and Slim Ouni

LORIA/INRIA
615 rue du jardin botanique 54602 Villers-lès-Nancy FRANCE


ABSTRACT

Our acoustic to articulatory inversion method exploits an original articulatory table structured in the form of a hypercube hierarchy. The articulatory space is decomposed into regions where the articulatory-to-acoustic mapping is linear. Each region is represented by a hypercube. The inversion procedure retrieves articulatory vectors corresponding to an acoustic entry from the hypercube table. A dynamic procedure is used to recover the best articulatory trajectory according to a minimum articulatory effort criterion. The inversion ensures that inverse trajectory non-articulatory parameters generate original formant trajectories with a very good precision, but not that they are realistic from a phonetic point of view. This work shows how additional simple articulatory constraints can be incorporated in the inversion process. Constraints are implemented in the form of bonus attached to the points which verify the constraints imposed. This enables the inversion to be guided towards more realistic inverse articulatory trajectories.

1. INTRODUCTION

Most of the acoustic-to-articulatory methods rest on an analysis-by-synthesis approach. That means that the articulatory-to-acoustic mapping is represented, either explicitly, in the form of a table giving acoustical parameters (in general, formants frequencies) for a number of points which sample the articulatory space [4], or implicitly, in the form of neural networks, for instance. The quality of the representation influences strongly inverse solutions recovered as these trajectories use points of the articulatory table. For these reasons, we have developed an adaptive sampling algorithm to ensure that the acoustical resolution is almost independent of the region under consideration in the articulatory space [8]. The adaptive sampling leads to a table that is organized in the form of a hierarchy of hypercubes, and ensures that, inside each hypercube, the articulatory-to-acoustic mapping can be approximated by means of a linear transform.

In our case we chose Maeda’s articulatory model [6] which describes the vocal tract shape with seven parameters expressed by means of a linear transform. The singular value decomposition [3] of the Jacobian matrix provides all the solutions of this equation in the form of a particular solution and the basis vectors of the null space of the Jacobian. As the dimension of vector $F$ is 3 and that of $P - P_0$ is 7, the dimension of the null space is generally 4.

In order to determine all the points of the hypercube which can produce measured formants, the intersection of the hypercubes with the space defined by the particular solution and the basis of the null space has to be sampled.

Let $P_{\text{mod}}$ be the particular point given by the SVD method, and $\{v_j\}_{j=1..4}$ the basis of the null space, solution points $P_s$ are therefore:

$$P_s = P_{\text{mod}} + \sum_{j=1}^{4} \beta_j v_j$$  \hspace{1cm} (1)

Coordinates $\beta_{j=1..4}$ have to be chosen to ensure that $P_s \in H_c$, i.e.:

$$\alpha_{in.f}^i \leq P_{\text{mod}}^i + \sum_{j=1}^{4} \beta_j v_j^i \leq \alpha_{sup}^i \quad i = 1..7$$  \hspace{1cm} (2)

where $\alpha_{in.f}^i$ and $\alpha_{sup}^i$ define the lower and the higher value of the $i^{th}$ articulatory parameter in the hypercube under consideration.

This problem is straightforward in dimension two but it did not received any solution in the general case. We therefore solved it in a approximate manner by considering two sets of linear programs, one to find lower value for each of the $\beta_j$s and the second to find the higher values of the $\beta_j$s. As the dimension of the null...
space is 4 generally, four linear programs are solved by using the simplex method. The extreme values of \( \beta_s \)'s define a domain that contains that of the solutions to the initial problem. It is therefore necessary to sample this domain to find points which belong to the hypercube. From a practical point of view we adapt the sampling step to keep a reasonable number of points (less than 100).

The frequency resolution imposed during the construction of the articulatory table enables an average precision of 10Hz between formants of the original signal and those synthesized from inverse articulatory points.

3. CONSTRUCTION OF THE INVERSE ARTICULATORY TRAJECTORIES

For each formant 3-tuple the previous procedure gives the set of articulatory points that are acceptable at one instant. Only one point among these points has to be kept to define an articulatory trajectory. The search of such a trajectory is achieved by means of a dynamic programming algorithm which minimizes articulatory efforts of the speaker. It has to be noted that the criterion used for minimization incorporates only the articulatory aspect since articulatory points recovered at the previous step ensure a good frequency accuracy (less than 10 Hz in average) between formants measured from speech and those obtained from inverse points.

Let \( s(i) \) be the set of articulatory points recovered at time \( t_i \). The sequence of these sets over the time interval of the signal to be inverted is : \( S = (s(0) \ldots s(i) \ldots s(N)) \) where \( N \) is the number of instants where inversion has been carried out. The construction of a trajectory gives rise to a double selection : instants \( i \) where the trajectory is defined and the articulatory point selected from the set \( s(i) \). This double selection allows all the articulatory points recovered at one instant where inversion failed to be eliminated. The selection of instants corresponds to a monotonic function \( j \). The resulting sequence is :

\[
\overline{S} = (s(j(0)) \ldots s(j(k)) \ldots s(j(K))) \quad \text{where} \quad K < N.
\]

The choice of one point among points selected gives the articulatory trajectory : \( \overline{\alpha} = (\alpha(j(0)) \ldots \alpha(j(k)) \ldots \alpha(j(K))) \) where \( \alpha(.) \) is the vector of the seven articulatory parameters at instant \( t_i \).

The cost of incorporating \( \alpha(j(k)) \) after \( \alpha(j(k-1)) \) is :

\[
C(\alpha(j(k)), \alpha(j(k-1))) = \lambda \sum_{i=1}^{7} m_i (\alpha_i(j(k)) - \alpha_i(j(k-1)))^2 + \text{Bon}(\alpha(j(k)))
\]

where \( m_i \) is the weight given to the \( i^{th} \) articulatory parameter and \( \text{Bon}(\alpha(j(k))) \) is a bonus that represents the interest of accepting the point \( \alpha(j(k)) \) in the final trajectory. The bonus must be greater than the first term of \( C(\alpha(j(k)), \alpha(j(k-1))) \) to prevent the algorithm from finding an empty trajectory. More importantly, it allows constraints to be imposed on the trajectories one wants to favor.

Therefore, the criterion minimized by the dynamic programming algorithm is :

\[
D = \sum C(\alpha(j(k)), \alpha(j(k-1)))
\]

Once the best trajectory has been found, it is regularized by means of an algorithm that improves the regularity while ensuring that formant trajectories obtained by synthesis from inverse articulatory trajectories are close to those measured in the original speech signal. The strength of this algorithm [5] is to take into account trajectories globally and to incorporate the acoustic behavior of the articulatory model.

4. INTRODUCTION OF CONSTRAINTS ON INVERSE TRAJECTORIES

Fig. 1. Formant trajectories for the first three formants (for transition /yi/) obtained by synthesis from inverse articulatory parameters without any constraint imposed. Solid lines represent synthesized trajectories. Original formant values are represented by points.

Fig. 2. Temporal evolution of three articulatory parameters (jaw, tongue position and protrusion) without any constraint imposed.

Inversion results for a large number of vowel-to-vowel transitions and /NCV/ are presented in [8]. Maeda’s model has been adapted to the subject who uttered the speech signals by means of the method proposed by Galvan [2]. The articulatory table was adapted to the subject who uttered the speech signals by means of the method proposed by Galvan [2]. The articulatory table was adapted to the subject who uttered the speech signals by means of the method proposed by Galvan [2]. The articulatory table was adapted to the subject who uttered the speech signals by means of the method proposed by Galvan [2]. The articulatory table was adapted to the subject who uttered the speech signals by means of the method proposed by Galvan [2]. The articulatory table was adapted to the subject who uttered the speech signals by means of the method proposed by Galvan [2].
The inversion procedure described above leads to articulatory trajectories of Fig. 2 (For sake of clarity, only the trajectories of the jaw, tongue position and protrusion are represented).

The most salient characteristic of these results is the weak protrusion of /y/ although it is one of the essential articulatory characteristic of /y/, even if re-synthesized formant trajectories are very close to the original ones (see Fig. 1).

This means that the criterion accepted to find the best articulatory path - minimization of articulatory efforts - is not sufficient without additional constraints. Rather than modifying it, we placed a constraint on the first point of the protrusion trajectory by setting it to $2.7\pm 0.1\sigma$. The incorporation of constraints in the dynamic programming algorithm amounts to give a very large bonus to starting points whose protrusion parameter is $2.7\pm 0.1\sigma$. It can be noted that it is very simple to impose constraints at other instants of the inversion by specifying a very large bonus for points verifying these constraints.

The new inverse trajectory for the protrusion parameter is more conform to what is expected, but it is nevertheless not sufficient, and above all, the starting point of the jaw corresponds to too open a position.

We therefore supplemented the constraint placed upon the first point by setting the position of the jaw to a fairly high value $1.5\pm 0.5\sigma$. Inversion results are now conform with those expected since protrusion decreases strongly from /y/ to /i/ and the closeness of the mount increases slightly without notable evolution of the tongue position (cf. Fig. 4 and 3). It is very important to mention that all these trajectories enable the formant transitions of the original signal to be re-synthesized with a high precision (cf. Fig. 5).

This example shows that the incorporation of very simple constraints allows the inversion to be guided towards expected solutions. Nevertheless, it should be noted that the number of so-
olutions satisfying constraints strongly decreases as their strictness increases. So, we were obliged to give a margin of error set to 0.5σ in order to get one solution at least. This means that we reached the acoustical compensatory limits of the articulatory model adapted to our subject, or equivalently, that the adaptation of the articulatory model is not sufficient to faithfully reproduce speech of our subject. During preliminary experiments dedicated to the adaptation of Maeda’s model we have used MRI images [7]. The adapted model was probably more faithful to the vocal tract geometry of the subject under consideration, than the model used in this inversion experiment, since the adaptation proposed by Galvan only involves acoustical data. However, the average error on the frequencies of the first three formants was fairly high (49 Hz for F1, 125 Hz for F2 and 170 Hz for F3). These errors are comparable with those observed by Story et al. [9], for instance. They stem from errors on the vocal tract geometry and physical parameters used in the computation of losses without being able to evaluate their relative contribution.

One of the advantages of our method is that it ensures that all the possible inversion solutions, given the articulatory model and the frequency precision set for formants recovering, can be explored. To the best of our knowledge, this is the only inversion method based on an articulatory model that may guarantee that all the trajectories allowed by the model are explored. With regard to inversion of the /yi/ transition presented above, one can note that there is only a small number of solutions verifying a constraint placed on a single point of articulatory trajectories. This confirms the relevancy of the articulatory model which, from a very limited constraint, enables deformations of the vocal tract, fitting well real observations, to be generated.

Exploring the space of possible articulatory trajectories quasi exhaustively leads to a high complexity as there are 100000 inverse points at each instant in average. The complexity of the standard dynamic programming algorithm would require $N \times 10^6$ computation of local connection costs approximately, where $N$ is the number of instants where inversion is carried out. As we allow possibly incomplete trajectories, with gaps of limited length (at most 3 points), this leads to an approximative complexity of $N \times 4 \times 10^3$. This important complexity is partly due to the exploration of the null space even if the choice of vectors in this space does not influence acoustical parameters.

On the contrary, gradient vectors define the acoustical behavior of the articulatory model and define efficient articulatory commands. We thus developed a multi-scale exploration strategy to reduce the complexity drastically. It consists of delaying the null space exploration after the determination of the overall articulatory strategy by relaxing the regularity constraint. The criterion which is minimized during the first step is a mixture of the articulatory strategy measured by cosines of gradient vectors and articulatory distance. This strategy reduces the complexity by a factor of 400 in average, and produces inverse solution only slightly different of those recovered by the standard algorithm.

5. CONCLUSION

The weak point of inversion methods using an articulatory table built from a partial sampling of the articulatory space is to orient inversion towards articulatory trajectories that rely only on points of the table. Consequently, some articulatory trajectories which are quite relevant from an articulatory point of view are not found.

On the contrary, our inversion method does not implicitly favor any articulatory solution, and the inversion example presented above shows that some inverse solutions do not respect articulatory characteristics observed in real speech. We have the project of supplementing the articulatory table by giving hypercubes the probability of being visited during the production of speech by real speakers. Learning these probabilities can be achieved through two methods. The first one consists of exploiting classical articulatory knowledge in the form of constraints to recover expected trajectories for a given utterance. The example above gives an idea of this process. By imposing protrusion and jaw position inverse trajectories are correct from a phonetic point of view this enables the probability of the corresponding articulatory points to be increased. This solution is somewhat similar to that proposed by Bailly [1] and requires the definition of a sufficiently large number of utterances for which articulatory gestures can be predicted correctly. The second solution consists of using true speakers to realize this learning by exploiting X-ray data or data obtained by electromagnetography.

6. REFERENCES