Can we retrieve vocal tract dynamics that produced speech?
Toward a speaker articulatory strategy model

Slim Ouni

LORIA
BP 239 - 54506 Vandoeuvre-ls-Nancy Cedex, France
First.Last@loria.fr

Abstract

In this paper, we argue the possibility of retrieving the vocal tract dynamics from speech as was produced by a given speaker. First, we present an inversion method that provides a complete set of articulatory solutions without excessive constraints. The selection of articulatory trajectories is not an easy task, as many trajectories are possible. Based on some inversion experiments and data observation, we show the importance of the introduction of the notion of speaker articulatory strategy to be able to model this variability.

1. Introduction

Estimating vocal tract shape from speech signal has received considerable attention because it offers new perspectives for speech processing. Recovering the vocal tract shape would enable knowing how a speech signal has been articulated. This potential knowledge could give rise to a number of breakthroughs in automatic speech processing. For instance, the location of critical articulators could be exploited to discard some acoustic hypotheses, in the case of automatic speech recognition. For language learning this could offer articulatory feedbacks. Lastly, in the domain of phonetics, inversion would enable knowing how sounds were articulated without requiring medical imaging or other measurement techniques.

Most of the acoustic-to-articulatory methods rely on an analysis-by-synthesis approach. That means the articulatory-to-acoustic mapping is represented, either explicitly, by pairs stored in a codebook for a number of points which sample the articulatory space, or implicitly, by neural networks, for instance. The quality of the representation influences strongly the recovered solutions as these trajectories use points of the codebook.

The acoustic-to-articulatory inversion is a difficult problem mainly because of the non-uniqueness of the relationship between the articulatory and acoustic spaces and the non-linearity of this relationship. To resolve this problem, we developed an inversion method that provides a complete description of the possible solutions without excessive constraints [1]. We presented an adaptive sampling algorithm to ensure that the acoustical resolution is almost independent of the region under consideration in the articulatory space. This leads to a codebook that is organized as a hierarchy of hypercubes, and ensures that, within each hypercube, the articulatory-to-acoustic mapping can be approximated by means of a linear transformation. The inversion procedure retrieves articulatory vectors corresponding to acoustic entries from the hypercube codebook.

At this point, we retrieve hundreds (and sometimes thousands) of solutions for each speech segment. The next step was to find articulatory trajectories that were the best match to the real speaker behavior, i.e., to retrieve articulatory trajectories that were very similar to the speaker vocal tract dynamics.

In this paper, we start by giving an overview of our inversion method using hypercubic codebook. Then, we discuss the issue of finding the best articulatory trajectories from the inversion solutions. First, we present our method used for this step. We show in particular that several choices are possible. Then we discuss how important to take into account inter-speaker and intra-speaker variabilities to find these trajectories.

2. Retrieving vocal tract articulations from speech

In order to find all the vocal tract shapes (represented by a set of articulatory parameters of Maeda’s model [2] used in our work) which can give rise a 3-tuple of formants representing a speech signal, we searched for all the hypercubes, whose acoustic image contains the formant vector. Then, for each of these hypercubes all the possible articulatory vectors were retrieved. Let \( F \) be the first three formant vector, \( F_0 \) the formant vector at the center of the hypercube and \( \nabla F \) the Jacobian of the articulatory-to-acoustic mapping computed at the center of the hypercube \( F_0 \). One searches for all the points \( P \) such that:

\[
F = F_0 + \nabla F(P - F_0)
\]

The singular value decomposition of the Jacobian matrix provides all the solutions of this equation in the form of a particular solution and the base vectors of the null space of the Jacobian. As the dimension of vector \( F \) is 3 and that of \( P - F_0 \) 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 hypercube with the space defined by the particular solution and the base of the null space has to be sampled.

Let \( P_{\text{svd}} \) be the particular point given by the SVD method, and \( \{v_j\}_{j=1...4} \) the base of the null space, solution points \( P_\delta \) are therefore:

\[
P_\delta = P_{\text{svd}} + \sum_{j=1}^{4} \beta_j v_j
\]  

Coordinates \( \beta_{j=1...4} \) have to be chosen to ensure that \( P_\delta \in H_c \), i.e.:

\[
\alpha_{\text{min}}^i \leq P_{\text{svd}}^i + \sum_{j=1}^{4} \beta_j a_j^i \leq \alpha_{\text{max}}^i \quad i = 1...7
\]

where \( \alpha_{\text{min}}^i \) and \( \alpha_{\text{max}}^i \) define the lower and the higher value of the \( i^{th} \) articulatory parameter in the hypercube under consideration.
We solved (2) in an approximate manner by considering two sets of linear programs, one to find lower value for each of the \( \beta_3 \)’s and the second to find the higher values of the \( \beta_3 \)’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 \( \beta_3 \)’s define a domain containing solutions of the initial problem. Therefore we sample this domain to find points in the hypercube. In practice, 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 codebook enabled a precision of 10 Hz in average between formants of the original signal and those synthesized from inverse articulatory vectors.

As shown in [1], this inversion method provides a quasi-exhaustive description of all the possible vocal tract shapes that give a 3-tuple of formants, and with a very little error on formant frequencies. This method thus enables the investigation of articulatory constraints that can be added to guide inversion and to recover realistic articulatory trajectories. The key point is that this inversion method enables a clear separation between the representation of the articulatory space and the incorporation of constraints or knowledge to guide inversion.

3. Recovering articulatory trajectories

In earlier experiments, the inversion method presented above gave between 500 and 8000 solutions for each speech segment. At this point, one should choose a set of articulatory trajectories according to some criteria.

As the solution space is potentially vast and solutions possibly not realistic from a phonetic point of view, constraints can be incorporated to focus the exploration in articulatory regions of interest. Schoentgen and Ciocea [3] introduced a local constraint based on either kinematic or potential energy to select one solution at each step of the inversion. We might also consider as choice criteria the smoothness and slow variation in time of the articulatory trajectories, as this is the general behavior of the vocal tract. We presented in [1, 4] a two step method for retrieving such trajectories.

The first step is a non-linear smoothing algorithm initially proposed in [5] which is a variation of the dynamic programming technique. In this algorithm, a weight could be assigned for each articulator and also a bonus could be used to force the choice of a particular inversion solution. This non-linear smoothing algorithm found smooth articulatory trajectories from the knowledge of inversion solutions recovered for each speech segment. The second step consists of regularizing the trajectories built by using the non-linear smoothing algorithm. This regularization is achieved through a variational method [1].

Even though, this method provides quite realistic articulatory trajectories, we have no evidence that such trajectories are actually produced by the speaker or can really exist. All what one can say is that the trajectories are smooth, and vary slowly. The examination of the vocal tract dynamics did not show any unnatural movement of any articulator.

Unfortunately, we have more evidence that finding the vocal tract dynamics that were those which produced speech are very difficult to retrieve. This statement is based on two facts: (1) some inversion experiments where we applied constraints on the inversion solution; (2) some observation of human speakers’ behavior.

3.1. Applying constraints to inversion solutions

In [6], we showed three examples of possible articulatory trajectories that we can retrieve by imposing each time some constraints. We showed, for instance, that we can impose different weights to some articulatory parameters during the smoothing step. The results were rather different behaviors as we obtained different trajectories, but were likely to be realistic. For example, when we imposed an important weight to the jaw, the trajectory of the jaw varied very slowly, and the other articulators get themselves organized to compensate for this situation, i.e. the non-linear smoothing algorithm searched among other solutions to keep the main criteria of smoothness valid.

In [4], we studied the solution space for the /yi/ transition as the main articulatory difference between /yl/ and /l/ is the protrusion which is strong for the French /yl/. All the recovered solutions provided a very good fitting between original and resynthesized formants and the main constriction and opening were correct. However, the best solution, in the sense of the criterion used for optimization in non-linear smoothing algorithm, did not present strong protrusion (see Fig. 1). In order to explore the solution space, we added a simple constraint on the lip protrusion (supplemented by a secondary constraint on jaw position). This constraint was implemented by adding a strong bonus, attached to the first point of the inversion. Fig. 2 shows that this simple constraint enables the recovery of a protrusion more conform to phonetic knowledge.

Figure 1: Dynamics of three articulatory parameters (jaw, tongue position and protrusion) without any imposed constraint [4].

Figure 2: Dynamics of three articulatory parameters (jaw, tongue position and protrusion) when imposing the protrusion to be close to 2.7\( \sigma \) and the jaw position to 1.5\( \sigma \) for the first point [4].
These two sets of experiments showed that we still have freedom in finding different trajectories, as the inversion method did not add any particular constraint while searching for all the solutions. By adjusting the smoothing algorithm parameters, we were able to obtain different trajectories. As for the /yi/ transition example, we showed that we were able to guide the inversion to the expected solution. It would be better if the guidance were automatic.

3.2. Observation of 3D-data

It is important to investigate the vocal tract behavior of real speakers while uttering speech, when such data are available. Probably having X-ray films is the best way to study vocal tract dynamics. Unfortunately, these data are very limited and not available for many speakers. We were able to get some 3D-data of some parts of the face (mouth opening, mouth protrusion and jaw height). These data were collected by Wrobel and colleagues [7] within the framework of developing a talking head and studying speech coarticulation. They used a stereovision technique to reconstruct 3d markers from two camera recordings. As a result, one can get the trajectories of some points of the lips (lip height and lip protrusion, for example) and the trajectory of the jaw height. One set of data was a recording of ten speakers uttering several VCCV and few sentences (see [7]).

We extracted from this corpus the trajectories of ten speakers uttering the French word “le joaillier” [la.zxɛ̃.je]. The speakers were 5 female and 5 male native speakers of French. In Fig. 3 and Fig. 4, we present the trajectories of the height of the lips, i.e., the distance between a marker on the upper lip and a marker on the lower lip (which we considered as an approximation of the mouth opening). As we can see in these figures, we have variabilities among the speakers. Each speaker used different strategy for uttering the same word. Four speakers presented similar strategies (SPF4, SPF5, SPM2 and SPM3). However, the six other speakers showed different strategies. Thus, this shows that to some extent, articulatory strategies are personal choices rather than imposed by physiological constraints (which is still true for some sounds, but other sounds have more articulatory freedom).

4. Which trajectories to choose?

As we saw in the previous section, it is unlikely to retrieve articulatory trajectories that produced a given utterance by a given speaker based solely on physiological constraints and even phonetic constraints. Speech production is planned at physiological level and at higher level (cognitive and phonological-phonetic processing). Several works showed that variability inter-speakers and intra-speakers does exist [8, 9, 10]. In addition to anatomical reasons, these variabilities can be explained by the fact that speaking rate induces to a reorganization of articulators to produce speech at the right rate. The context and speech segments interaction can also explain this variability. Finally, these choices are simply personal preferences, habits or social conventions [10].

So, which trajectories should one choose? The answer depends on the purpose of the inversion.

If the purpose is to reproduce articulatory trajectories that are recognizable by human, i.e., lead to intelligible visual speech, probably it is not important to find the same trajectories as those produced by the speaker who uttered speech signal. However, the result should be correctly recognized by human, i.e., does not present any vocal tract shape not usually used by a particular community of speakers. For this purpose, perceptual evaluation is needed to valid the articulatory production. A talking head might be a good tool to show vocal tract dynamics to participants in perceptual experiments, and they have to evaluate the articulation of the talking head. For this purpose, an important step is to drive the talking head by an articulatory synthesizer, which is not trivial, and need probably some approximations as we will have a mapping from 2D to 3D space. In this case, the parameters that need a fine precision will be the visible part of the vocal tract, i.e., lips and jaw. We would have probably more freedom to choose different trajectories for the tongue and larynx, for example.

If the purpose is to retrieve exactly what the speaker did while uttering speech signal, we need to take into account the personal choices of the speaker, as mentioned earlier. We need to know these choices during trajectory recovering step. Thus, before choosing an articulatory solution vector, we need to check whether this particular speaker is very likely to choose it.

Figure 3: Mouth opening dynamics for 5 female speakers uttering the French word “le joaillier” [la.zxɛ̃.je]. Time stamps are in milliseconds. The vertical axis is the distance between two markers on the lips.
This is what we call the speaker articulatory strategy. It is represented by a model incorporating these choices. Thus, it is speaker dependent. This model includes a description of the different phoneme realizations articulately and modeling the coarticulation. In fact, it is well known that coarticulation is an important phenomenon in speech as it describes the interaction between neighboring speech segments in the articulatory space. Thus an articulatory speaker strategy model is to some extent based on a coarticulation model (as in [11] for instance). Practically, this model would be used during the selection process of the solutions to build the articulatory trajectories. There are two ways to train the speaker strategy model: (1) heuristically, defining manually the model based on some expertise; or (2) automatically, using an automatic training method.

Choosing the second solution would be the best. For this purpose, a set of data would be collected and used for training. We started the collection and analysis of the data (as presented in [7]). This allows us to collect measurement of visible part of the vocal tract. Some complex measurements might be needed for the inner part of the vocal tract, but it is not easy to collect for new speakers. Defining the speaker strategy model on some articulators would reduce the number of allowable trajectories by the inversion method.

Our future work will be an explicit definition of this model and its training. The model will combine a coarticulation scheme and an optimization criterion as presented in [1] specific to a given speaker. The validation will be the comparison between the collected data and the obtained trajectories.

5. Acknowledgements

The author thanks B. Wrobel-Dautcourt, M.O. Berger, Y. Laprie and all participants in the project LABIAO for providing the 3D data.

6. References