Inversion experiments based on a descriptive articulatory model

Yves Laprie, Slim Ouni, Blaise Potard, Shinji Maeda

INVERSION EXPERIMENTS BASED ON A DESCRIPTIVE ARTICULATORY MODEL

Yves Laprie [1], Slim Ouni [2], Blaise Potard [1] & Shinji Maeda [3]

[1]Speech group, LORIA/CNRS, Nancy France
[2] Perceptual Science Laboratory, University of California - Santa Cruz
[3] CNRS-URA 820 and ENST-Dept. TSI, Paris, France

ABSTRACT: Our goal is to recover articulatory information from the speech signal by acoustic-to-articulatory inversion. Like most inversion methods proposed in the literature, our method relies on the analysis-by-synthesis paradigm. After an overall description of the inversion method the paper presents how inversion can be used to investigate acoustic properties of the articulatory model, which helps us to formulate the way of incorporating effective constraints to obtain phonetically realistic inverse solutions. First, the inversion method has been applied to French vowels in order to study the dispersion of vocal tract shapes corresponding to each of these vowels. Second, in addition to increasing the intrinsic quality of the articulatory table, constraints can be incorporated directly in the dynamic programming stage to obtain more realistic inverse trajectories in a sequence of vowels. These constraints derive from phonetic knowledge, such as the observation of the strong protrusion for /y/, of a posterior place of articulation for /u/, of a strong mouth opening for /a/ and so on.

INTRODUCTION

Our goal is to recover articulatory information from the speech signal by acoustic-to-articulatory inversion. Like most inversion methods proposed before, our method relies on the analysis-by-synthesis paradigm that exploits an articulatory model. The choice of the vocal tract representation is crucial since it determines the number of parameters to recover. As acoustic data often consist of sets of the first three formant frequencies, a description of the vocal tract as concise as possible must be adopted to reduce the under-determination of the problem. Besides, as shown by Atal (Atal 1978) a 3-tuple of formants can be generated by an infinity of vocal tract shapes even if losses are taken into account in the acoustical simulation. On the other hand, the phonetic exploitation of articulatory models goes against this idea of models describing the vocal tract with a very small number of parameters. Indeed, one generally prefers vocal tract shapes that can be easily interpreted from a phonetic point of view, which is not possible with very rough vocal tract approximations.

The existing solutions may be classified in terms of conciseness and anthropomorphism. The first family is that of area function models (Stevens 1955, Fant 2001…). Their main weakness is the difficulty to introduce constraints intended to guarantee that vocal tract shapes may be realized by a human speaker. Conversely, nothing guarantees that these models are powerful enough to represent all the possible vocal tract shapes a human speaker can realize. Increasing their flexibility would lead to substantially increase their number of parameters, and consequently weaken their essential strong point. Their conciseness highly facilitates the inversion; and sometimes this allows the inverse problem to be expressed in a closed form. This is the case of methods derived from linear prediction and methods using local physical constraints (Schoentgen 1995). Beyond the fact that area function models do not represent the vocal tract faithfully, their phonetic interpretability and adaptability to a new speaker are low.

The second family is that of models intended to approximate the sagittal slice of the vocal tract. The oldest models are purely geometrical (Mermelstein 1973) and the most recent ones rely on the processing of vocal tract images (either X-ray images for (Maeda 1979) or MRI for (Engwall 1999)). The strength of these models is their faithfulness with respect to the vocal tract geometry. Most often these models define the sagittal shape from linear components obtained by analyzing a large number of images. Unlike purely geometrical models, articulatory parameters correspond to deformations of the vocal tract produced by true speakers. Consequently, these models correctly cover the domain of

vocal tract shapes that a human speaker can produce with a relatively small number of parameters - between 7 and 9. It is likely that the number of parameters of an area function model should be increased twofold to cover vocal tract shapes with the same faithfulness. However, except for true 3D models as those of Badin and Engwall, the third dimension (section areas) must be approximated from the knowledge of the sagittal slice that is only a 2D information. Another strong point of sagittal slice models is the possibility to adapt models to a new speaker by modifying pharynx and mouth sizes.

The counterpart is linked to the double nature of the under-determination inherent to these models. For an area function model the multiplicity of inverse solutions only depends on acoustical equations, whereas for articulatory models this multiplicity also depends on the fact that several articulatory parameters configurations can give approximately the same vocal tract shape, and consequently the same area function. However, as our long term objective is to recover information that can be exploited from a phonetic point of view, we have, therefore, adopted Maeda's model.

DESCRIPTION OF OUR METHOD

Our inversion method exploits an articulatory table that associates vectors of articulatory parameters, i.e. 7-tuples in the case of our model, with their corresponding 3-tuples of the first three formant frequencies. This table represents the synthesis facet of the inversion. It is used to recover all the possible 7-tuples of articulatory parameters corresponding to the formant frequencies extracted from a vowel signal at each time sample. The second stage consists of reconstructing articulatory trajectories that are sufficiently regular along time. This is achieved by a dynamic programming algorithm that minimizes a cost function that represents the overall “distance” covered by articulatory parameters. The final stage is to improve the articulatory regularity and the acoustical proximity of formants derived from the determined model vocal tract to the original formants measured on the speech segment.

Construction of the articulatory table

The strength of our inverse method resides at the quasi-uniform acoustic resolution of the articulatory table. This property originates in the construction method that evaluates the linearity of the articulatory-to-acoustic mapping at each step. Articulatory parameters of Maeda's model vary between -3σ and 3σ where σ is the standard deviation. Thus, the codebook inscribes a root hypercube. Sampling the articulatory space amounts to finding reference points that limit linear regions. Therefore we evaluate linearity on segments between any two vertexes in the hypercube considered: acoustic values are linearly interpolated at the middle point between two vertexes from the acoustic values calculated at these vertexes and the result is compared against that directly given by the articulatory synthesizer. If the difference between the acoustic values synthesized and those interpolated is less than a predefined threshold, the hypercube is considered to be linear concerning the articulatory-to-acoustic mapping. Otherwise this hypercube is decomposed into sub-hypercubes and the linearity test is repeated for every new hypercube. This procedure is repeated recursively until the hypercube edge becomes smaller than a predefined value or no non-linearity higher than the predefined threshold exists anymore.

Improving the construction of the articulatory table

Actually, the cube decomposition is also applied if one or several vertexes of the cube lead to a non valid vowel area function, i.e. an area function with a complete constriction. Such cubes delimitate the articulatory space. As for linearity, the decomposition is stopped if the edge becomes smaller than a predefined value and invalid cubes at the boundary are discarded. A small edge value was thus used to guarantee that the boundary of the articulatory space is known with a sufficient precision. Unfortunately, it turned out that the boundary decomposition led to a vast number of small cubes since each decomposition gives 128 (2^7) new cubes even if only a very small number of vertexes among them are invalid. Conversely, accepting a coarse edge to discard cubes with some invalid vertexes would lead to the elimination of articulatory regions that are important to produce cardinal vowels and particularly /u/. Therefore we decided to decompose cubes with invalid vertexes only if the ratio of invalid vertexes with respect to the total number of vertexes, i.e. 128, is greater than a predefined threshold, or if the jacobian at the hypercube centre cannot be calculated.

Exploration of the null space of the articulatory-to-acoustic mapping

For every acoustic entry given by the first three formants, the inversion process consists in finding all the possible hypercubes, i.e. those for which the articulatory-to-acoustic mapping can produce the 3-tuple of formants observed. Then for each of these hypercubes precise solutions have to be found. As the inversion consists of recovering seven parameters from the first three formants, the solution space has four degrees of freedom. This means that the null space of the local articulatory-to-acoustic mapping (i.e., the inverse mapping to zero of linear application corresponding to the Jacobian matrix of the hypercube) has to be sampled to get a good description of the solution space. This is not a trivial problem since it corresponds to find the intersection of a 4-dimensional space and a 7-dimensional hypercube. A first approximation of the intersection is obtained by linear programming. Then the belonging of each articulatory sample to the hypercube is tested (Ouni 2001).

INVERSION EXPERIMENTS

Finding all the places of articulation for isolated vowels

Using real data, X-ray or MRI images for instance, is the best way to study the place of articulation for vowels. However, this approach suffers from several limitations due either to the hazardous nature of the imaging techniques or to the non-natural recording conditions. Even though Wood's work was considered the most complete study of the place of articulation for vowels using real data, the quantity of the data was still limited and thus prevented all possible articulatory configurations to be covered.

A solution to study vowels with respect to the place of articulation is to recover articulation information directly from the speech signal. The advantage is to enable a large amount of vowel examples to be analysed. Our inversion method guarantees that all the possible articulatory configurations, i.e. inverse solutions for a given vowel represented by its 3-tuple of formants are found (in relation to the sampling step chosen to explore the null-space). Therefore, we exploited it to study the articulation place for French vowels uttered by a male speaker (Ouni 2003).

Figure 1 shows the inverse solutions for the vowel /u/ in three different representations: lip area (Al) with respect to the main constriction area (Ac), Al with respect to the main constriction position (Xc), and Ac with respect to Xc. Xc varies between 0 (glottis) and 17.5 cm (lips). Each cross represents {Xc, Ac, and Al} for one articulatory solution. It turns out that there are precise Xc that correspond fairly well to those expected even if a number of possible vocal tract shapes present constriction areas substantially larger that those observed by (Wood 1979) for comparable vowels. For example, there are approximately 80% of the inverse solutions for /a/ that present Ac larger than 2.5 cm^2. In the case of the /u/ however, there are almost no solutions with Ac larger than 2.5 cm^2. Three main Xc can be identified for the /u/: one around 11 cm (palatal constriction), one around 6 cm (velar) and one around 3.5 cm (pharyngeal). We can note than 88 % of the solutions have a palatal constriction, whereas only 3% are pharyngeal. In French, only the palatal /u/ is used, but in other languages a velar /u/ can also be observed. The pharyngeal /u/, however, has never been observed. This example shows how using phonetic knowledge can help improve the inversion process: measuring the main constriction area would offer a mean to eliminate, or at least, to penalize unrealistic vocal tract shapes (e.g. the shapes corresponding to pharyngeal /u/) in an articulatory codebook.

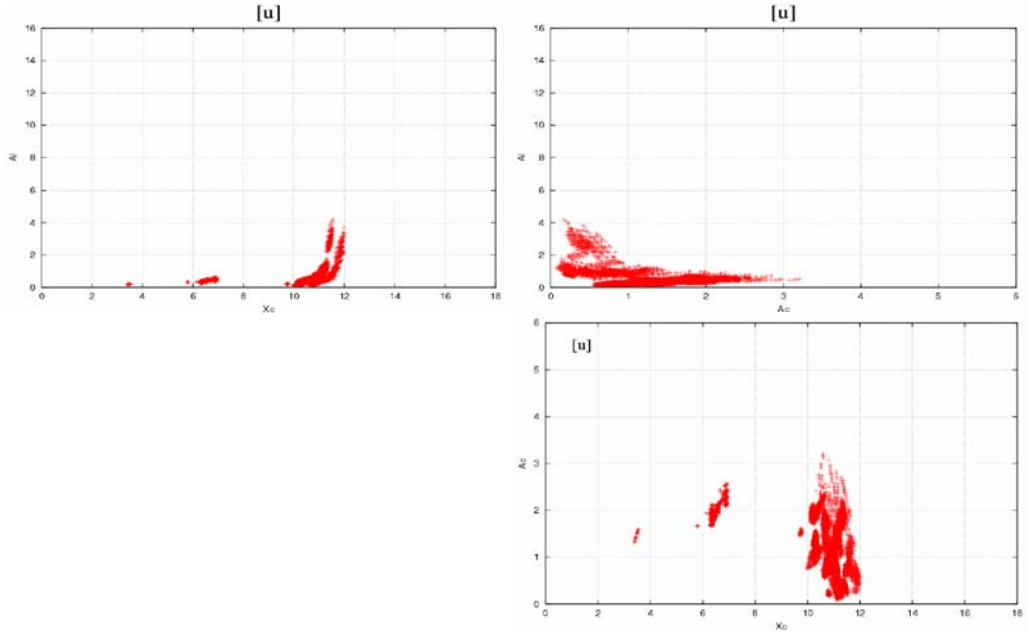

Figure 1. Inverse solutions for the vowel /u/ in the planes Xc/Ac, Xc/Al and Ac/Al, where Xc is the position of the main constriction measured as the distance from the glottis along the length of the vocal tract in cm, Ac the constriction area in cm^2, and Al the lip opening area in cm^2.

Adding phonetic constraints to improve the realism of inverse solutions

The inverse method presented above can give a possibly vast number (i.e. 10,000) of vocal tract shapes represented by a 7-tuple of articulatory parameters for a given 3-tuple of first three formant frequencies. Therefore, a minimal cost algorithm is exploited to build a smooth articulatory trajectory from vocal tract shapes recovered at each time sample. The criterion minimized mainly corresponds to the distance covered by articulatory parameters and with no information about the muscular energy available, only articulator inertia or acoustical efficiency is taken into account. It should be noted that all the potential inverse articulatory trajectories give re-synthesized formants very close to the original ones even if they are not realistic from a phonetic point of view.

One of our objectives is to investigate constraints that can be used to improve the phonetic realism of articulatory trajectories recovered. Indeed, there is an important number of inverse articulatory trajectories that cannot be realized by a human speaker.

Constraints can be provided by either general phonetic knowledge on vowel articulation or the observation of visible articulatory parameters (jaw and lip parameters) as it is the case when one can see the speaker. The advantage of the first solution is that it can be incorporated in the articulatory codebook eventually.

In a first time we incorporated phonetic knowledge directly in the dynamic programming stage to obtain more realistic inverse trajectories. As we observed that the inverse trajectory recovered for /yi/ does not preserve the protrusion of /y/ we added a constraint on protrusion. This amounts to give a very large bonus to the first point of the trajectory that, consequently, will be preferred by the dynamic programming minimization. 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.

Figure 2 represents the sequence of vocal tract shapes recovered for /yi/ after the introduction of constraints placed on the protrusion and jaw position of the initial point of the vowel /y/. This solution respects both tongue position and protrusion that strongly decreases from /y/ to /i/.

Improving the acoustic faithfulness of Maeda's model

The articulatory model was built by applying a data analysis method to a series of X-ray images for a female speaker uttering 10 short sentences. As the acoustic signal has been recorded during shooting, these data are very interesting to assess the relevance of the inversion method. However, even if the overall resynthesized formant trajectories are quite similar to those extracted from speech, we observed a non-negligible deviation between original and resynthesized formants. Actually, the model geometrical precision depended on two scale factors that have been set arbitrarily because the calibration of the X-ray machine was not known precisely. The adjustment of these scale factors was not possible in 1979 when the model was constructed because it would have required too much computation time. Therefore, we sampled ranges of reasonable values for these two factors to find better values. It turned out that the ad-hoc area increase set to 40% in the original model can be removed provided than the scaling factor is set to 196 (instead of 187). The overall frequency error thus decreases from 114Hz to 54Hz for formants F1 and F2.

CONCLUSION AND FUTURE WORKS

The strength of our inverse method resides at the quasi-uniform acoustic resolution of the articulatory table together with the exploration of the null space of the articulatory-to-acoustic mapping. This guarantees that almost all the inverse solutions can be explored given a predefined articulatory sampling step.

As can be observed in inversion experiments a vast number of articulatory trajectories can be recovered. Some of them originate in compensatory effects that are actually exploited by human speakers and some originate in the under-specification of the articulatory model. Our objective is to control the incorporation of constraint precisely to investigate their merit. This is possible because our inversion method is quite neutral unlike some other methods that implicitly favour solutions but with a penalty to others that could be realistic.

Evaluating inversion is difficult because there are few data associating acoustic and articulatory data together. Therefore, we are now exploiting data used to build Maeda's articulatory model. This will enable the investigation of the precision required for model adaptation and the acoustic precision required to recover correct articulatory gestures as well.

Furthermore, we are now working on the incorporation of static and dynamic constraints into the inversion process. Static constraints can be derived from phonetic features for speech sounds and directly used to penalize articulatory codebook entries that do not satisfy articulatory features expected for one 3-tuple of formants. Dynamic constraints could be statistically learnt from articulatory data. However, there are probably not enough data to realize a correct learning. Therefore, we prefer to investigate how constraints derived from the observation of visible articulatory parameters can be exploited.

REFERENCES

Atal, B. S., Chang, J. J., Mathews, M. V. & Tukey J. W. (1978), "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", Journal of Acoustical Society of America **63**:5 1535-1555

Engwall, O. (1999), "Modelling of the vocal tract in three dimensions", Proceedings of the 6th European Conference on Speech Communication and Technology, Budapest, **1** 113-116

Fant, G., (2001), "Swedish vowels and a new three-parameter model", TMH-QPSR **1** 43-49

Laprie, Y. & Ouni , S. (2002) "Introduction of constraints in an acoustic-to-articulatory inversion method based on a hypercubic articulatory table " Proceedings of the International Conference on Spoken Language Processing, Denver

Maeda, S. (1979) "Un modèle articulatoire de la langue avec des composantes linéaires", *Actes des 10èmes Journées d'Etude sur la parole*, 152-162.

Mermelstein, P. (1973), "Articulatory Model for the Study of Speech Production", Journal of Acoustical Society of America **53** 1070-1082

Ouni, S. & Laprie, Y. (2001), "Exploring the Null Space of the Acoustic-to-Articulatory Inversion Using a Hypercube Codebook", Proceedings of the 7th European Conference on Speech Communication and Technology, Aalborg, **1** 277-280

Ouni, S. & Laprie, Y (2003), "A study of the main constriction of the vocal tract for French vowels using an acoustic-to-articulatory inversion method", Proceedings of the International Congress of Phonetic Sciences, Barcelona

Schoentgen, J. & Ciocea (1995), S., "Direct calculation of the vocal tract area function from measured formant frequencies", Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, **1** 745-748

Stevens, K. N., & House, A. S. (1955), "Development of a quantitative description of vowel articulation", Journal of Acoustical Society of America **27** 484-493

Wood, S. (1979), "A radiographic analysis of constriction locations for vowels", Journal of Phonetics, **7** 25-43

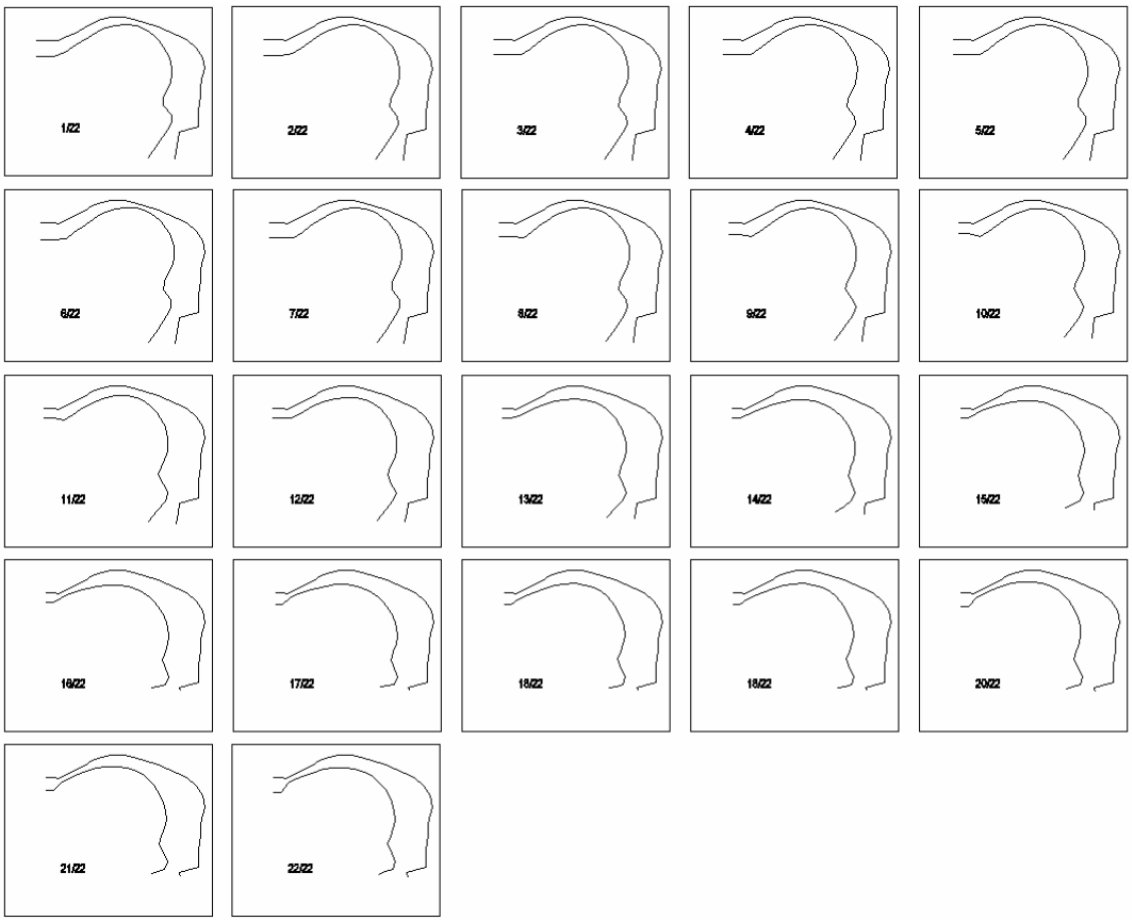

Figure 2. Sequence of inverse solutions for /yi/ after the introduction of constraints on protrusion