Using phonetic constraints in acoustic-to-articulatory inversion

Blaise Potard, Yves Laprie

Using phonetic constraints in acoustic-to-articulatory inversion

Blaise Potard, Yves Laprie

LORIA - Campus Scientifique
BP 239 - 54506 Vandoeuvre-les-Nancy Cedex
potard@loria.fr

Abstract

The goal of this work is to recover articulatory information from the speech signal by acoustic-to-articulatory inversion. One of the main difficulties with inversion is that the problem is underdetermined and inversion methods generally offer no guarantee on the phonetical realism of the inverse solutions. A way to adress this issue is to use additional phonetic constraints.

Knowledge of the phonetic caracteristics of French vowels enable the derivation of reasonable articulatory domains in the space of Maeda parameters: given the formants frequencies (F1,F2,F3) of a speech sample, and thus the vowel identity, an "ideal" articulatory domain can be derived. The space of formants frequencies is partitioned into vowels, using either speaker-specific data or generic information on formants. Then, to each articulatory vector can be associated a phonetic score varying with the distance to the "ideal domain" associated with the corresponding vowel.

Inversion experiments were conducted on isolated vowels and vowel-to-vowel transitions. Articulatory parameters were compared with those obtained without using these constraints and those measured from X-ray data.

1. Introduction

Atal and his colleagues[1] have shown that an infinity of area functions can give exactly the same 3-tuple of formants. One of the challenges in acoustic-to-articulatory inversion is thus to add constraints which reduce the number of inverse solutions without eliminating relevant solutions. One common approach is to use an articulatory model that generates only relevant vocal tract shapes. These 2D or 3D models are generally derived from medical images acquired for one subject by applying some factor analysis technique. Even if an articulatory model substantially reduces the range of possible vocal tract shapes there still exists a very large number of inverse solutions for each 3-tuple of formants.

Actually, it turns out that the articulatory variability is one of the essential characteristics of speech production. The articulators of speech have large compensation capacities that enable the production of one sound one after the other even if its intrinsic articulatory characteristics are very far from those of the other. Despite this large variability there exist a number of expected articulatory invariants. The aim of the work reported in this paper is to exploit standard phonetic knowledge to express these articulatory invariants in the form of constraints imposed to articulatory parameters.

Other classes of constrains have been investigated. Physiological constraints, for instance, give ranges of possible articulatory parameters and/or constraints about the maximal acceleration or jerk (third derivative of position) acceptable for speech. However, most of these constraints require the knowledge of parameters that cannot be easily accessible. The main advantage of phonetic constraints is that they can be easily expressed and that they present a great robustness with respect to speaker variability.

At first we describe the phonetic constraints and their implementation in our acoustic-to-articulatory framework[2], which uses an articulary table (or codebook), generated using Maeda's articulatory model[3]. Then we evaluate them in the case of isolated vowels to investigate their effects in terms of place and degree of constriction, and in the case of speech utterances for which the articulatory parameters are known.

2. Phonetic features as articulatory constraints

The main idea behind the use of phonetic constraints is the assumption that each phoneme has invariant articulatory features, like a strong protrusion for the french /y/, for instance. In the case of vowels, which present slow time varying acoustic structures in comparison to other phonemes as stop consonants, these features can be easily translated into constraints on the articulatory parameters.

2.1. Phonetic constraints for vowels

In the particular case of vowels, four types of constraint can be defined : the mouth opening, the protrusion of the lips, the lip stretching, and the position of the tongue dorsum. The relevance of each constraint depends on the vowel considered. As mentioned in the introduction there exists a strong inter-speaker variability. We thus designed numerical, rather than boolean, constraints that return a phonetic relevancy from the knowledge of formants.

Tab. 1 summarizes our classification for the 10 non-nasals French vowels. D stands for "tongue dorsum position", O for "mouth opening", S for "lip stretching", and P for "lip protrusion". The convention we use for classification is straightforward: the higher the number, the higher the value associated with the given constraint. For example, a constraint O_1 means that the mouth has a small opening, a value of O_4 means a very big opening. These data are average values of the way native French speakers articulate vowels, and thus may be different from the way a particular speaker articulates sounds of French. Note that for the main place of articulation of vowels, corresponding to D in the case of vowels, the range of possible values is a sub-domain of the values acceptable for consonants (from 0 for /p,b,m/ to 9 for /ʁ, ɹ/). This explains why D ranges only between 6 and 8 for vowels.

Table 1: *French vowels classification.*

Vowel	D	O	S	P
i	D6	01	S4	P1
e	D6	02	S3	P1
ɛ	D6	03	S2	P1
a	D7	04	S1	P1
y	D6	01	S1	P4
ø	D6	02	S1	P3
œ	D6	03	S1	P2
u	D8	01	S1	P4
o	D8	02	S1	P3
ɔ	D8	03	S1	P2

2.2. Transposing phonetic constraints in the articulatory model

In most articulatory models, transposing simple phonetic features into parameters of the model can be quite complex. In our case, we use Maeda's model[3], in which the parameters can be easily interpretable from a phonetic point of view. Consequently, expressing phonetic constraints in terms of articulatory parameters is straightforward: lip protrusion and tongue dorsum position are already parameters of the model, and the mouth opening is a linear combination of two parameters (jaw position, and intrinsic lip opening).

Actually, this constraint also uses the tongue position in order to take into account compensatory effects described in [4]: Maeda observed that for non-rounded vowels (/i/, /a/, /e/), the tongue position and the jaw opening had parallel effects on the acoustic image, and therefore were mutually compensating. He also observed that this compensatory effect was indeed used by his test subjects. Furthermore, it appeared that the direction of compensation did not depend on the vowel pronounced: there was a linear correlation

$$Tp + \alpha Jw = \text{ Constant}$$

, where Tp is the tongue position, Jw the jaw position, and the α the linearity coefficient that is the same for both /a/ and /i/. The other vowels were not studied because there were not enough occurrences of them in the X-ray database. Maeda observed this compensation in both his subjects (but the coefficients of correlation were of course different). The coefficient we used for PB was the one Maeda found experimentally on X-ray data, which was approximately equal to 0.66. This compensatory effect allowed Maeda to explain most of the articulatory variability for /a/ and /i/.

2.3. Acoustic space partitionning

For each phoneme, we have to define an acoustic domain where the phonetic constraints are considered to be valid, that is, a domain where we are likely to observe articulatory configurations which respect the given constraints. We could compute these domains directly from the articulatory model, by synthesizing the domains of the phonetic constraints: in future works, we may use self-organising maps like Kohonen's. But currently, we use simple models, centered on the average vowels formant frequencies of French speakers.

Currently, our model works on the 3-D space of the first three formant frequencies. We tested different models for the partitionning of the acoustic space : Voronoi diagram around the vowels (cf. Fig. 1); Voronoi diagram weighted by the standard deviation of each formant frequencies (cf. Fig. 2.

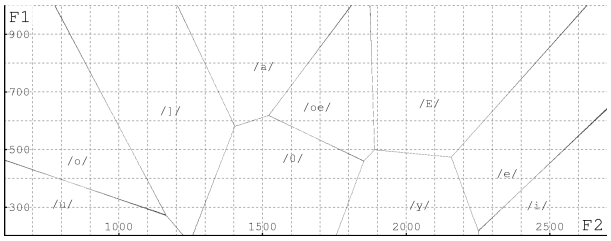

Figure 1: *Voronoi diagram model.*

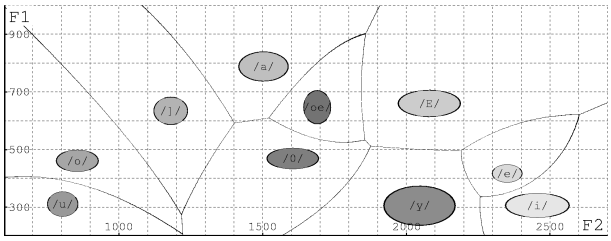

Figure 2: *Ponderated Voronoi diagram model.*

2.4. Phonetic scoring

Now that we have partionned the acoustic space, we still have to explain how a phonetic score can be associated to each inverse solution: basically, a given acoustic vector is attached to an "ideal articulatory domain", as defined by the constraints in Tab. 1, corresponding to the region of the acoustic space it belongs to. Then each inverse solution V corresponding to this 3-tuple can be given a "phonetic score", according to the distance of the articulatory vector to the "ideal domain". A simple way to do that would be to compute the norm of the vector defined by the point and its orthogonal projection onto the domain. Actually, we compute a score relative to each type of constraint: tongue dorsum, mouth opening, lip stretching and protrusion.

The computation of the score depends on two values: the target value of the constraint considered $\theta(v,t)$, where v is the vowel, and t is the type of constraint considered, and a margin $\sigma(v,t)$, which defines a validity interval $I(v,t) = [\theta(v,t) - \sigma(v,t); \theta(v,t) + \sigma(v,t)]$. If the value of the constraint for V is within $I(v,t)$, then it gets a perfect score (1) for that type of constraint. Otherwise, it gets a positive score less than 1 which exponentially decreases from 1 according to the distance to $I(v,t)$. The overall phonetic score is simply a linear combination of the 4 types of constraints, to get scores within the interval $[0;1]$ (1 being the best score). In our current model, all constraints have equal weight, except for the lip stretching which has a null weight, because Maeda's model cannot account for lip stretching, since it was designed using X-ray images of sagittal profiles of the vocal tract.

3. Experiments

We conducted inversion experiments on the original data Maeda used for his model. It consisted in a corpus of 10 sentences for a total time of about 20 seconds of X-ray cineradiography. Cardinal vowels and some VV sequences were selected in the speech signal, the first three formants frequencies were manually ex-

tracted. We built a high precision codebook adapted to Maeda's speaker. Although we studied the original speaker used to build the articulatory model, we still had to adapt the model to improve the acoustic faithfulness[2] because the geometrical calibration of the X-ray acquisition is not known precisely.

Despite this adaptation it must be kept in mind that the articulatory model together with the acoustic simulation are not capable of generating formant frequencies that have been measured from the original speech signal. Even by using articulatory parameters measured from X-ray images and the best geometrical adaptation the average error on F1 is still 54 Hz. This non negligible discrepancy is explained by the approximation of the recovery of the 3D information (corresponding to the area function) from the 2D information (corresponding to the sagittal profile of the vocal tract) provided by the articulatory model. This approximation, based on the method proposed by Heinz and Stevens[5], is unable to render the area everywhere from the glottis to lips precisely. In addition, physical constants involved in the acoustic simulation probably introduce a slight error. In conclusion, despite this favourable situation (the speech signal to be inverted has been pronounced by the speaker whose X-ray data have been processed to derive the articulatory model) the inversion is non trivial and cannot precisely recover the original articulatory trajectories.

3.1. Codebook caracteristics

Tab. 2 summarises the characteristics of the codebook used for inversion. The first line gives the number of unique articulary vectors which acoustic image was calculated during the codebook construction. The second line gives the number of linear hypecubes which were kept in the codebook. The third line gives the total[1] number of vertexes of the forementioned hypercubes. The fourth line gives the percentage of the total volume of hypercubes of the codebook over the whole articulatory space explored. The fifth line gives the maximum (over the first three formant frequencies) average absolute error of the formant frequencies linearly interpolated from the codebook data over the formants computed using the articulatory model. The acoustic precision used in the codebook construction for the linearity test was 0.3 bark on each formant frequency.

Table 2: *Codebook characteristics.*

Number of points sampled	607,422,368
Number of hypercubes	1,071,353
Number of vertexes	137,133,184
Articulatory space kept	32.9 %
Average acoustic precision	8.3 Hz

3.2. Checking the model consistency

As the phonetic constraints, as well as the acoustic space partitionning, are independant of the speaker in our current model, we beforehand checked that the acoustic domains correspond to the images of the phonetic constraints domains. For each vowel, we plotted the acoustic images of articulatory vectors that had perfect phonetic scores, and we could observe that for each vowel, the acoutic domain was included in the overall acoustic image of the corresponding "ideal" articulatory domain. We also computed a new partition of the acoutic space by attributing each point of the acoutic space to the vowel that had most images in its neighborhood (each vowel had the same number of synthesized articulatory points, randomly chosen in the ideal domain). The resulting F1/F2 graph was very close to our acoustic models.

[1] the actual number of unique articulatory vectors is lower than this number, which is simply the number of hypercubes multipled by the number of vertexes in an hypercube, that is, $2^7 = 128$.

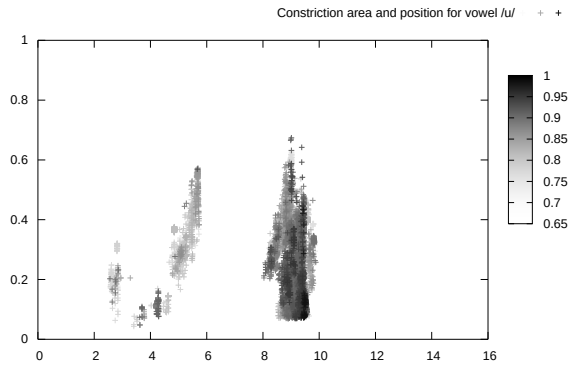

Figure 3: *Phonetic scoring of the inverse solutions for /u/*

3.3. Inversion of isolated vowels

Vowels /a/,/i/,/u/,/e/, /o/, were inverted using the phonetic constraints. The inverted points are each given a phonetic score varying with their distance from the "ideal domain". Fig. 3 represents the area at the maximum constriction (in cm^2) as a function of its position (in cm, starting from the glottis) for each inverse solution found. The gray level of each point is a function of its phonetic score, darker points have a higher score. Although constraints are applied on articulatory parameters, they give rise to a consistent overall effect, i.e. they enhance the emergence of well located regions in the plane spanned by the place of maximal constriction and the constriction area, and weakens some secondary places of articulation. These regions are furthermore more consistent with the articulatory data of Wood[6]. The second observation is that these phonetic constraints penalize vocal tract shapes with large constriction areas. This aspect is important because the acoustic properties of vocal tract shapes are not very sensitive to a general and uniform area increasement. This thus enables this kind of unrealistic vocal tract shapes to be penalized.

3.4. Inversion of VV sequences

We extracted several VV sequences from the sentences uttered by PB: /ui/, /yi/, /ie/.

Since the audio signal was quite noisy, we had to extract formants by hand. After that, the sequences were inverted using different kinds of constraint. Here, we present the results for the sequence /yi/. For all the figures the time unit is *ms* and the articulatory parameters are given in standard deviation with respect to the neutral position.

Fig. 4 represents the 3 main parameters (jaw, tongue position, lip protrusion) as measured on the X-ray images.

Fig. 5 is the inverted sequence using only biodynamic constraints on the articulatory parameters: that is, the "overall velocity" of articulators is minimized. Although the inverse so-

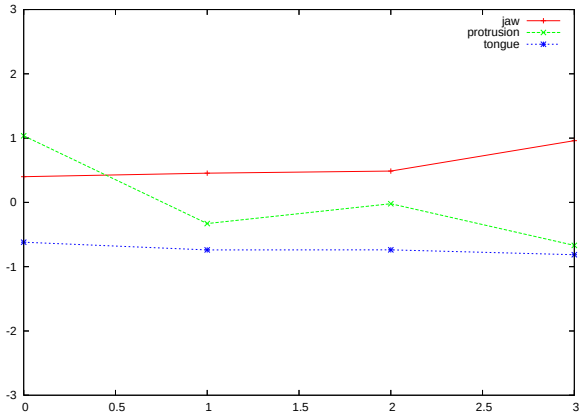

Figure 4: *Measured articulatory parameters*

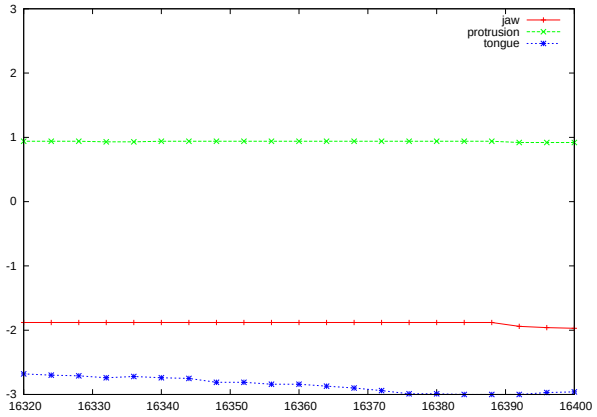

Figure 5: *Inversion with biodynamic constraints only*

lution has a very good acoustic precision, it is very different from the observed solution, and it is not phonetically realistic. Not surprinsingly, the minimization of the overall velocity gives rise to quasi-straight transitions.

Fig. 6 is the inverted sequence, using both biodynamic and phonetic constraints, with equal weights. It should be noted that the original trajectories are sampled at a lower rate (50 Hz) than the inverse trajectories. This time, the solution is much more realist. The overall articulatory movements have been recovered properly even if absolute values of the articulatory parameters are not equal to the original ones. As mentioned above this is due to the acoustic mismatch between the articulatory acoustic simulation and the human process of speech production.

This experiment shows that very general constraints, derived from standard phonetic knowledge, enable the recovery of realistic articulatory trajectories. The impact of these phonetic constraints is all the more sensitive since our inversion method exploits a quasi exhaustive description of the articulatory space.

4. Conclusion and perspectives

The under determination of the acoustic-to-articulatory problem has given rise to several directions of research in order to incorporate constraints that can compensate for the lack of data. However, most of the constraints envisaged (see [7] for instance) require the knowledge of numerical constants difficult to be estimated. In comparison with these constraints pho-

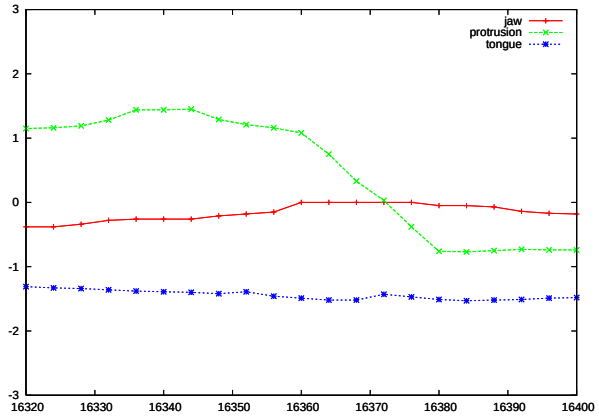

Figure 6: *Inversion with phonetic and biodynamic constraints*

netic constraints present two advantages. Firstly, they do not involve numerous numerical parameters, which is a key point. Secondly, they are very general, speaker independent and have been extensively validated since they derive from standard phonetic knowledge. Furthermore, these phonetic constraints could be easily coupled with constraints derived from the observation of the speaker's face.

Acknowledgments

We would like to thank Dr. Shinji Maeda fruitful discussions and for making his articulatory model and data available.

5. References

[1] 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," *JASA*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[2] Y. Laprie, S. Ouni, B. Potard, and S. Maeda, "Inversion experiments based on a descriptive articulatory model," in *6th International Seminar on Speech Production*, Sydney, Autralia, Dec. 2003.

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

[4] ——, "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. J. Hardcastle and A. Marschal, Eds. Kluwer Academic Publishers, 1990.

[5] J. M. Heinz and K. N. Stevens, "On the relations between lateral cineradiographs, area functions and acoustic spectra of speech," in *Proceedings of the 5th International Congress on Acoustics*, 1965, p. A44.

[6] S. Wood, "A radiographic analysis of constriction locations for vowels," *Journal of Phonetics*, vol. 7, pp. 25–43, 1979.

[7] 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.