Adapting visual data to a linear articulatory model

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

[1]LORIA-CNRS 615 rue du jardin botanique
54600 Villers-lès-Nancy FRANCE

potard@loria.fr, laprie@loria.fr

Abstract. *The goal of this work is to investigate audiovisual-to-articulatory inversion. It is well established that acoustic-to-articulatory inversion is an under-determined problem. On the other hand, there is strong evidence that human speakers/listeners exploit the multimodality of speech, and more particularly the articulatory cues: the view of visible articulators, i.e. jaw and lips, improves speech intelligibility. It is thus interesting to add constraints provided by the direct visual observation of the speaker's face. Visible data was obtained by stereo-vision and enable the 3D recovery of jaw and lips movements. These data were processed to fit the nature of parameters of Maeda's articulatory model. Inversion experiments were conducted.*

1. Introduction

The main difficulty in acoustic-to-articulatory inversion is that there is no one-to-one mapping between the acoustic and articulatory domains. Furthermore, the problem is under-determined, as there are more unknowns that need to be determined than input data available and therefore there is a large number of vocal tract shapes that can produce the same speech spectrum. One important issue is thus to add constraints that are both sufficiently restrictive and realistic from a phonetic point of view, in order to eliminate false solutions.

Speech is a bimodal signal which comprises the acoustic signal and the view of the speaker. These two modalities are strongly correlated and redundant. There is strong evidence that human speakers/listeners exploit the multimodality of speech, and more particularly the articulatory cues: the view of visible articulators, i.e. jaw and lips, improves speech intelligibility in adverse conditions (Sumby and Pollack, 1954; Le Goff, 1997).

The aim of the work presented in this paper is to supplement acoustic data used in our acoustic-to-articulatory inversion framework (Ouni and Laprie, 2005; Potard and Laprie, 2005) by data provided by the view of visible articulators (lower jaw and lips). In this paper, we mainly present the method used to adapt the recorded 3D visual data to Maeda's articulatory model (Maeda, 1979).

2. Adaptation of visual data

2.1. Data acquisition

We exploited data acquired with a stereovision system designed to study labial coarticulation (Wrobel-Dautcourt et al., 2005; Robert et al., 2005). Our system only uses two fast

*This work is part of the ASPI project funded by the IST Programme of the Commission of the European Communities as project number IST-2005-021324.

Fig. 1. Stereo images of one speaker, EK, with 15 markers on the face.

B&W cameras (120 fps), a PC and painted markers that do not change speech articulation and provide a sufficiently fast acquisition rate to enable an efficient temporal tracking of 3D points. The corpus was mainly intended to study inter-speaker variability of labial coarticulation. 15 markers were painted on the speaker's face (only 4 markers on lips, Fig. 1) to keep the overall subject preparation time reasonable. In addition to markers used to study coarticulation or build a talking face we put 6 markers on the upper part of the face to compensate for the global motion of the head. The 3D positions of the markers were recovered using a stereo-vision procedure, as described in (Wrobel-Dautcourt et al., 2005).

2.2. Adapting visual data to Maeda's articulatory model

Maeda's articulatory model has been derived from X-ray sagittal images by applying a factor analysis (Maeda, 1979) which enables the explicit choice of linear components if needed.

It is the case of the jaw whose movements can be readily determined by measuring the position of incisors which appear very clearly on Xray images. Similarly, 3D face data enable the direct measurement of lip opening and stretching by measuring the position of markers painted on the lips (see Fig. 1). The protrusion is also estimated in the same manner. However, protrusion corresponds to a complex movement that implies some "unfolding" of the lips. The movements of markers painted on the lips in the sagittal plane thus only partially render this complex movement. Consequently, protrusion is probably slightly underestimated.

Unlike these articulatory parameters that can be directly derived from the measures, other parameters cannot be estimated directly and thus require some factor analysis. In the case of 3D face data the movement of the lower jaw (common to visible and articulatory data) cannot be measured directly from data. Indeed, the movement of markers painted on the chin depends on jaw movement but also on that of the lower lip which pulls these markers when it moves.

From the acquired visual data, we can compute four parameters: the mouth opening, lip stretching and jaw movement, which are straightforward to compute, and the lip protrusion, which is more complex:

- the mouth opening is simply the distance between the two sagittal points located on the upper and lower lips.
- the lip stretching is the distance between the two corner points of the mouth.
- the jaw movement is the distance from the points on the chin to a fixed point. We chose the averaged position of the four points located on the chin. By doing

that, we assume that the positions of the points of the chin only depend on the movements of the jaw, which, as said above, is not true, but in this case, we assume the influence of the lip movements to be negligible.

- the lip protrusion is the hardest parameter to compute. It is determined by using the projection of the sagittal points of the upper and lower lips, on a plane defined by the average positions of the four lips points.

The exploitation of these parameters derived from speaker face images requires that face and vocal tract articulatory parameters are consistent together. Indeed Maeda's model involves three parameters related to the speaker's face: jaw opening, lip opening and protrusion. The adaptation consists in expressing the visual parameters in the coordinate system of the vocal tract articulatory model. Two solutions can be envisaged.

The first consists in applying exactly the same factor analysis to visual data as that applied by Maeda to X-ray data. Then, articulatory parameters derived from the speaker's face are used in the same manner as the other articulatory parameters of Maeda's model. The underlying hypothesis is that both speakers, i.e. the one used to build the vocal tract articulatory model and the one whose face images are used during inversion, share common articulatory behaviour to prevent mismatches between the two models. This hypothesis is actually very strong.

The second solution consists in matching the visual measures obtained through acquisition to the dimensions of the corresponding visual features obtained through the articulatory model. The articulatory-to-visual relationship was inversed, and the discrepencies were accounted for by mapping the average values of the visual measures onto those of the X-ray data, and using contraints on the regularity of the articulatory parameters obtained. The expected advantage is to keep the internal consistency of the vocal tract articulatory model since there is no model for the face data. We now present these two methods.

2.3. Adapting the data through a factor analysis

Before describing the adaptation itself, let us describe the factor analysis used by Maeda for elaborating the articulatory model. The method, described precisely in Maeda (1990) consisted in an analysis in arbitrary factors. In the case of the lip parameters, the normalized parameters were simply decorrelated in a specific order to obtain orthogonal parameters: measurements of the jaw position, lip vertical opening, and lip protrusion and horizontal opening were thus used to obtain 3 orthogonal parameters (actually four, but only one factor was kept to express both lip protrusion and horizontal opening). The specific order in which the decorrelation occured was chosen by Maeda to take into account the specificity of his data. In particular, the position of the lower jaw was measured exactly, and thus was chosen as the main parameter. The influence of the jaw position on the other three data sets was removed by computing its correlation to each data set and subtracting it. The same process was then conducted using the decorrelated lip height on the remaining two data sets. Finally, a PCA was conducted on the last two data sets to obtain one last parameter, assimilated as the lip protrusion parameter. We should notice that Maeda implicitly assumed that the effective lip height was not influenced by the lip protrusion.

Our data is very different from Maeda's one. The first problem is due to the fact that in our case the jaw position is not known with a good precision. In particular, lip

movements make the skin of the chin slide, so the jaw parameter obtained in our measurements does not correspond to the intrinsic one. Furthermore, the "lip opening" data set does not really correspond to the actual lip opening either, since lip protrusion makes the lips unfold, and thus make sagittal points move relatively to the mouth boundary . Finally, our "lip stretching" data set, measured using markers on the mouth corners has almost nothing in common with what Maeda used (the actual lip horizontal opening). For that reason, we did not perform the PCA, and we used the uncorrelated lip protrusion parameter obtained at the second stage as the final parameter.

On top of these concerns with the data, there are issues regarding the method itself. The data measured has no particular reason to have the same standard deviation, or even average value, as the one used by Maeda. Using the parameters obtained using this method directly into the articulatory model might thus be problematic, since they would likely lead to incorrect articulators positions (except if the data sets had the same means and standard deviation as the original speaker, which would be very unlikely).

2.4. Direct adaptation using the articulatory model

The idea consists in using visual data directly as articulatory parameters. However, the visual face data does not represent exactly the X-ray measures used by Maeda. An adaptation is thus necessary.

Maeda's articulatory model computes the position of articulators from the articulatory parameters, and thus among others the position of the lips. Since the transformation is linear, it is reversible and it thus is easy to compute the articulatory parameters from the measured data. The main concern is that our data sets do not represent exactly the same measures as the one obtained through the articulatory model.

For all data sets, we have a common problem: the reference from which the data is measured is either unknown, or not relevant. We solve this by forcing the mean of each visual data set to coincide with the mean of its corresponding value from the X-ray data used in the construction of the articulatory model. This adjustment can be biased, since the means of these values may not be exactly equal in reality.

In addition, we also have to remove the influence of lip protrusion to obtain the lip opening from the distance of the sagittal points on the lips as explained in §2.2, and similarly the influence of lip opening to obtain the jaw position. Other possible sources of discrepency are negliged, in particular the influence of protrusion the markers of the chin.

To solve this second problem, the opening of the mouth is expressed as a linear combination of the distance between the sagittal points, and the protrusion parameter, this value being shifted so that its mean over the data of a given speaker is the same as Maeda's speaker. Likewise, the jaw position is being expressed as a linear combination of the measured value and the lip opening parameter. The optimal values for the linearity coefficients are found automatically by minimizing a criterion based on trajectory regularity and distance to neutral value, on the articulatory parameters obtained. For the sake of place, the precise computation cannot be extensively described here.

Before computing the linarity coeeficients, the articulatory model is adapted to each speaker using Galván-Rodrigez (1997) method. For the optimal parameters found, the final value of the lips opening was manually verified watching the corresponding

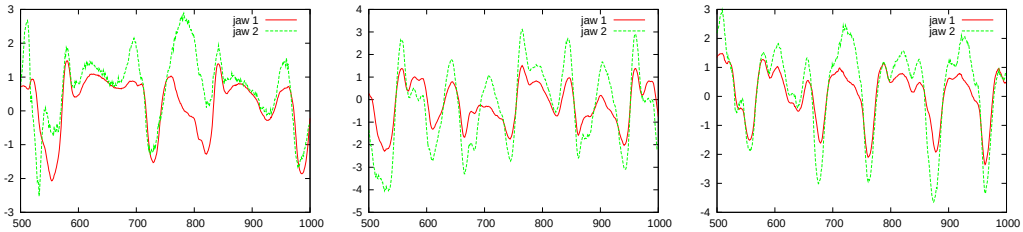

Fig. 2. Jaw parameters obtained respectively for speakers AB, BP, EK, using respectively the first method (plain line) or the second method (dashed line). Abscissa is the frame number, one frame being taken approximately every 8ms; ordinates are in standard deviations units (relative to the data set in the first case, relative to Maeda's model parameters in the second case).

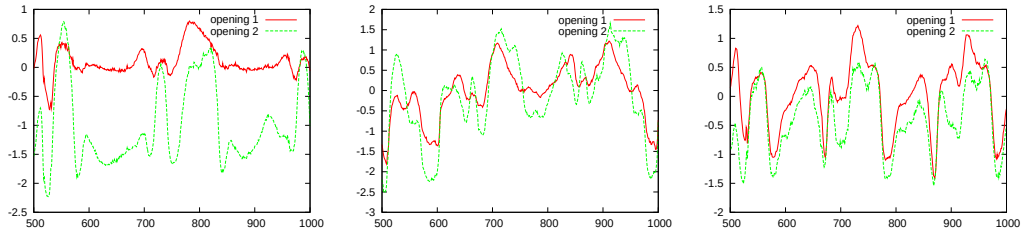

Fig. 3. Lip opening parameters obtained for speakers AB, BP, EK, using respectively the first method (plain line) or the second method (dashed line).

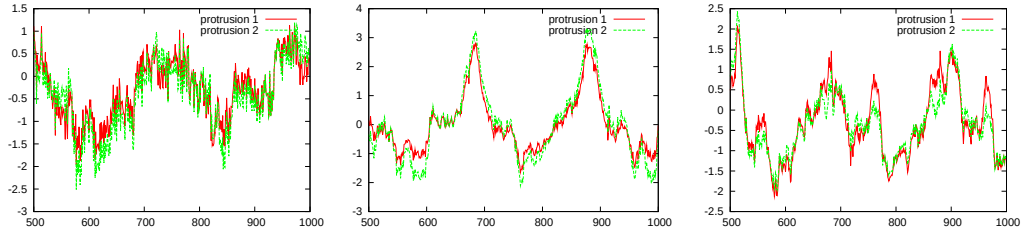

Fig. 4. Protrusion parameters obtained for speakers AB, BP, EK, using respectively the first method (plain line) or the second method (dashed line).

images. This second method is expected to give much better results than the first one.

3. Experiments

Several experiments were conducted to compare the articulatory parameters computed through these two methods. These experiments were conducted on three native French speakers: two males (BP and EK) and one female (AB). First, a comparison of the parameters obtained through the two different methods was conducted, and then some speech sequences were inverted using these parameters as additional constraints.

3.1. Comparisons of the two methods

Comparing the trajectories obtained using the two methods, we can learn interesting things on the behaviour of articulatory parameters. In the first method, we assumed that the articulatory movements of the speakers we study have the same statistical characteristics as the one used to build the model. In the second method, we directly compute the parameters that give the correct dimensions. By comparing the data obtained in each case, we can check the validity of our hypotheses.

On Fig. 2, the range of possible values (in standard deviation units, relative to the corresponding data set in the first case, relative to the corresponding parameter in the

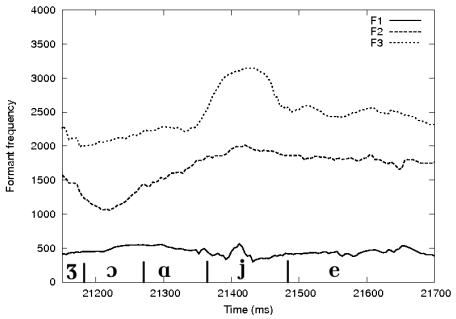

Fig. 5. Formants trajectories of the sample inversed, abscissa is the time in ms.

articulatory model in the second case) for the jaw parameter obtained through the second method is larger than what we usually allow in our inversion experiments, since it goes up to 4.5 (instead of 3) and down to -5.5 (instead of -3). Although this might seem extreme, it actually is not. The synthesized mouth shapes are visually similar to the original ones, and furthermore MaedaMaeda (1990) did mention that compared to other speakers, his model had small jaw movements; it thus is not surprising that other speakers would have larger variations for this parameter. We can see that for this parameter, the hypothesis we made with regards to the first method is clearly wrong.

The "lip opening" parameter (Fig. 3) has a range of values slightly smaller than Maeda's for EK, but slightly larger for AB and BP. We can also notice an important discrepancy between the values obtained through the two methods in the case of AB.

Finally, the lip protrusion (Fig. 4) is the most interesting parameter. For the three speakers, we obtain extremely close trajectories with the two methods. This result seems to indicate that the range of protrusion is fairly constant among native French speakers.

3.2. Inversion experiments

Using these parameters as additional input to inversion, we conducted inversion experiments on one sentence in our corpus, "Le joaillier a broyé les cailloux de la voyageuse," especially designed to evaluate inversion easily since most of the sounds are vowels, semivowels or other voiced sounds, using our inversion framework (Ouni and Laprie, 2005; Potard and Laprie, 2005), which uses an articulatory codebook of "linear" hypercubes.

We present one of these experiments in this article, in which as input, in addition to the three first formants frequencies, we use the parameters obtained using respectively the first and the second method. In these cases, we only perform the codebook inversion: that is, the articulatory trajectory displayed is a trajectory from the articulatory vectors selected from the codebook obtained through dynamic programming. The criterion used is simply articulatory movements minimization. Since the visual parameters are unreliable by nature, there is a relaxed selection applied (in this case, we retained all points generated from the codebook whose visual articulatory parameters are within an euclidian distance of 1 from the visual parameters targets, whereas we apply a strict selection –less than 3% error– on the formants frequencies).

The formants trajectories inversed are displayed on Fig. 5. Fig. 6 displays the results of inversion on "joaillier" /ʒɔɑje/ for method 1: we present the trajectories found

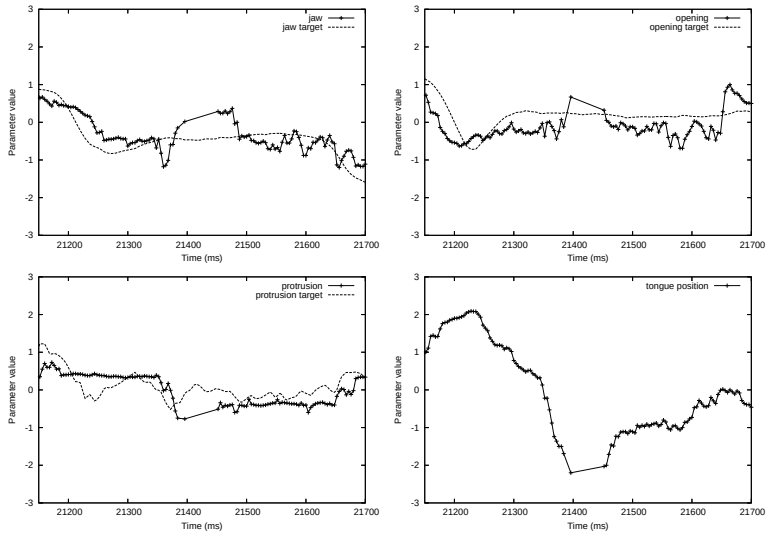

Fig. 6. Inversion results for “joaillier” using the first method; four articulatory parameters are displayed: jaw, lip protrusion, mouth opening and tongue position. When available, the corresponding input visual parameter is displayed in dashed ligne, whereas the inverse trajectory is displayed as crosses relied by segments. Each parameter can vary in interval [-3; 3], abscissa is the time in ms.

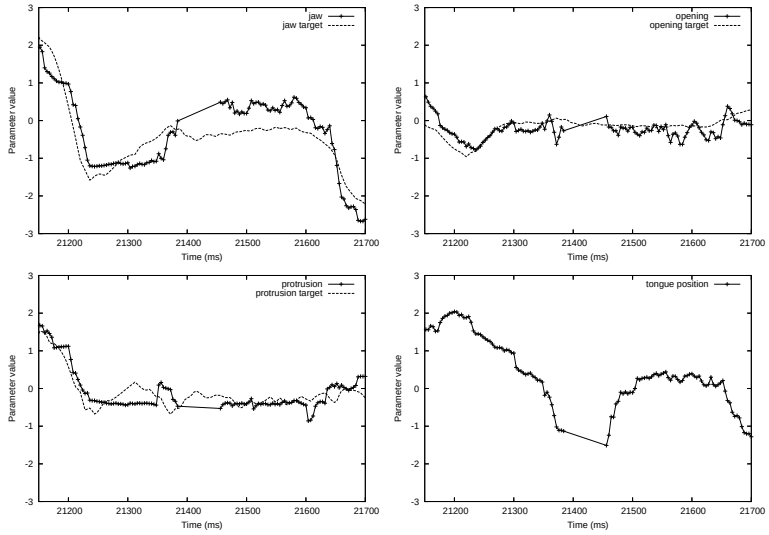

Fig. 7. Inversion results for “joaillier” using the second method; each parameter can vary in interval [-3; 3] and abscissa is the time in ms.

for the 4 main parameters (jaw, mouth opening, lip protrusion, tongue position). We also display in smooth line the respective visual constraints. As can be seen on the graph of the jaw, the inversion had trouble in the middle of the sequence, to render the /ɑj/ transition, since there is no solution found at all. The other two visual parameters are fairly close to their constraints. We can also observe that the tongue position trajectory is fairly consistent with what one would expect: it starts by getting in the back of the mouth to pronounce /ɔɑ/, then goes in front for /j/, and goes back for /e/ (when this parameter increases, it means the tongue gets further back).

The results of inversion using the second method (Fig. 7) are quite similar, although the visual constraints (and thus the articulatory trajectories recovered) are quite

different. The same behaviours are observed: the /ɑj/ transition is rendered with difficulty, the articulatory trajectories are very close from the observed values, and the tongue position parameter has a correct behaviour. This experiment is another illustration of the remarquable compensatory capabilities of Maeda's articulatory model, since even though the input visual data is fairly different, it still finds valid inverse solutions. Both visual data adaptation models appear equally efficient in this particular experiment.

4. Conclusions

This study of audio-visual inversion already shows very promising results. In particular, it appears that the model manages to find inverse solutions that respect the visual constraints, even when these constraints are distant from the "real" parameters values. Additionally, it seems that the method of acquisition used in our lab, coupled with our adaptation to Maeda's model can be a convenient way to study inter-speaker variability of articulation. This work will be pursued in several directions: we will conduct the same experiments on the other speakers of our corpus, to investigate if similar patterns can be observed (particularly the behaviour of the protrusion parameter). We will also investigate whether the use of additional constraints, such as phonetic constraints (as in Potard and Laprie (2005)) further improves the quality of inversion.

References

Galván-Rodrigez, A. *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives*. Thèse de l'Institut National Polytechnique de Grenoble, 1997.

Le Goff, B. Automatic modeling of coarticulation in text-to-visual speech synthesis. In *Eurospeech'97 Proceedings*, volume 3, pages 1667–1670, Rhodes, Greece, 1997. European Speech Communication Association.

Maeda, S. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

Maeda, S. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. and Marchal, A., editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

Ouni, S. and Laprie, Y. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *JASA*, 118(1):444–460, 2005.

Potard, B. and Laprie, Y. Using phonetic constraints in acoustic-to-aticulatory inversion. In *Interspeech, Lisboa*, pages 3217–3220, September 2005.

Robert, V., Wrobel-Dautcourt, B., Laprie, Y., and Bonneau, A. Strategies of labial coarticulation. In *Interspeech, Lisboa*, September 2005.

Sumby, W. H. and Pollack, I. Visual contribution to speech intelligibility in noise. *JASA*, 26(2):212–215, 1954.

Wrobel-Dautcourt, B., Berger, M. O., Potard, B., Laprie, Y., and Ouni, S. A low cost stereovision based system for acquisition of visible articulatory data. In *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05)*, pages 145–150, Vancouver, 2005.