Adaptation of Maeda's model for acoustic to articulatory inversion

B. Mathieu, Y. Laprie
CRIN-CNRS & INRIA Lorraine
BP 239 - 54506 Vandœuvre-lès-Nancy
e-mail: mathieu@loria.fr, laprie@loria.fr
FRANCE

Abstract

We are working on performing acoustic to articulatory inversion by using Maeda's model. The purpose of this work is to adapt the model to a new speaker. The adaptation quality is assessed by verifying that vowels uttered by the speaker lie inside the vocalic space defined by the model. It is with this aim in view that we realized a series of MR images for eleven oral French vowels (/i, e, ɛ, y, ∅, œ, a, ɑ, ɔ, o, u/). The adaptation may include modifications of: scale factors for the pharynx and the mouth cavity, the wall of the vocal tract and coefficients for the calculation of the area function from a sagittal shape. The scale factors have been determined by superimposing Maeda's model on the MR images. The wall has been obtained by calculating the mean value of the exterior contours of the vocal tract in the image series. As some discrepancies between natural and synthetic vowels remained the wall contour has been iteratively optimized by means of formant sensibility functions calculated for each section in the vocal tract. The inversion is carried out by means of a table-lookup procedure constrained by the smoothness of articulatory trajectories.

1 Introduction

We are working on performing acoustic to articulatory inversion by using Maeda's model [8]. This anthropomorphic model ensures that the vocal tract shapes generated are realistic from an articulatory point of view contrary to more geometrical models or even rough approximations by means of a small number of uniform tubes. This prevents the inversion process from incorporating very artificial and strong constraints on the evolution of vocal tract shapes. However, before any inversion procedure, the model must be adapted so to allow it to generate sounds a target speaker may generate. The purpose of this work is twofold: to adapt Maeda's model to make inversion possible and to investigate whether or not this model can be easily adapted to a new speaker. This study requires that MR (Magnetic Resonance) images are taken during sustained phonation of vowels, but in the future we will investigate how the adaptation could be achieved from the utterance of some vowels alone.

This will allow us to investigate the sensitivity of the inversion process with respect to the accuracy of the speaker adaptation and to compare inversion results with those obtained with rougher models.

2 Acquisition of images and speech signals

A General Electric SIGNA machine (1.5T) was used for the MR image acquisition. We accepted the parameters used by Yang et al.[12]. We realized a series of images for eleven oral French vowels (/i, e, ɛ, y, ∅, œ, a, ɑ, ɔ, o, u/). Three mid-sagittal slices with a 5 mm thickness are available for each vowel: one situated exactly at the mid-sagittal plane, one at 5 mm to the left and one at 5 mm at the right. The articulatory model was in fact obtained from X-ray images which project views of the whole head onto a single plane, whereas MR images correspond to a 5 mm slice. We accepted the image at 5 mm to the left of the mid-sagittal plane because it appears as the closest one to the X-ray images.

MR imaging has the advantage of giving good quality images without known bio-hazard. On the other hand, noise produced by the machine prevents recording of the speech signals and Lombard effect produces speaker articulation deformation.

Taking this into account we recorded the machine noise and instructed our male subject to pronounce the same vowels while listening the noise through a headphone. This allowed us to obtain vowels close to those the subject uttered in the MR machine.

3 Adaptation

Maeda's variable articulatory model allows the shape of the vocal tract to be calculated from seven parameters (jaw position, tongue location, tongue shape, lip aperture, lip protrusion and larynx height). In addition to the seven basis vectors corresponding to these parameters used for the calculation of the interior contour $i(x)$ of the vocal tract, the definition of the model also includes:

- two scale factors which control the sizes of pharynx and mouth cavities,
- the wall of the vocal tract (i.e. the exterior contour $e(x)$),
- coefficients α and β [6] which are used for the calculation of the area function given by $A(x) = \alpha(x)\,|i(x) - e(x)|^{\beta(x)}$ where x varies from 0 at the glottis to the length of the vocal tract at the lips.

The adaptation consists in modifying the two scale factors, followed by the exterior contour determination in two steps. The results are evaluated by measuring the error between formants of the natural vowels and of vowels synthesized by means of a frequency simulation from the articulatory model. Furthermore we verify that the natural vowels lie inside the vocalic space defined by the model. Before adaptation the mean error is 46 Hz for F1, 209 Hz for F2 and 184 Hz for F3; /u,o,ɔ,ɑ, a/ clearly lie outside the vocalic space.

The first stage computes the two scale factors, with the aim at making exterior contours extracted from MR images coincide with those produced by the model [9]. To achieve this goal we developed an image display package to analyze MR images from an articulatory point of view. The analysis process includes editing facilities to acquire vocal tract contours and a toolbox to deform Maeda's model. We determined the scale factors by superimposing the articulatory model on MR images. The scale factors accepted are those which lead to the best overall fitting between images and model.

For the pharynx and the mouth we obtained an expansion factor of 1.18 and 1.08 respectively. The fact that Maeda's model has been obtained for a female speaker explains these large values. We next determined the exterior contour e of the vocal tract by calculating the mean value of the exterior contours in the image series. We then extracted for every French oral vowel articulatory parameters that minimize the euclidian distance between the contour produced by the model and that of the corresponding image (Fig. 1).

The three first formant frequencies for every vowel are compared to verify the accuracy of the model, the mean error is: 48 Hz for F1, 136 Hz for F2 and 236 Hz for F3. A large error was observed for /ɑ/ which our speaker did not obviously pronounce well during image acquisition (the sound was strongly centralized). We also found that F3 is too weak for back vowels which is in agreement with results of Candille [4] and Story [11]. However, the adapted model better covers the vocalic space of our subject.

Some discrepancies observed between natural and synthesized vowels stem from the fact that images used for constructing Maeda's model are X-ray images. In particular, the exterior contour accepted is not well suited for Maeda's model. So starting from the contour extracted from MR image, we optimized the exterior contour ($e(j)$ where j is the section index in the tract) through an iterative process. At each step and for each vowel, the error on formant is reduced by modifying $e(j)$ according to the sensibility function given by the Jacobian matrix $\frac{\partial F_i}{\partial e(j)}$ (where i is the formant index). The exterior contour becomes the mean of contours obtained for each vowel and the process is repeated until the contour no longer varies.

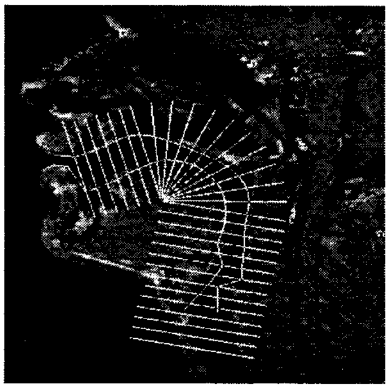

Figure 1: MR image of /ɔ/ and approximation of the vocal tract by Maeda's model adapted to our subject.

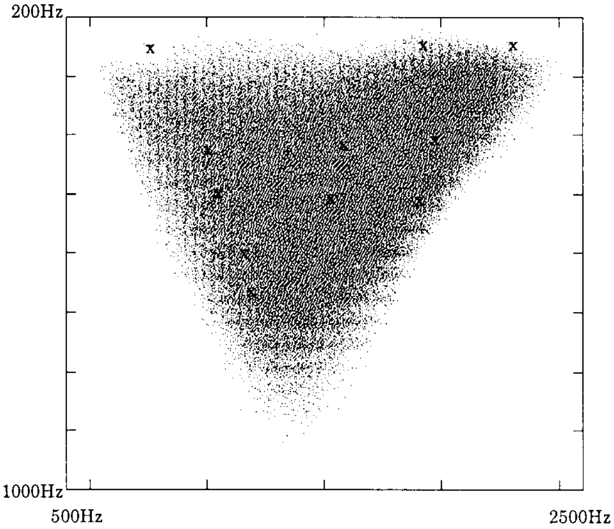

Figure 2: F1/F2 space (x points denote vowels uttered by the speaker)

In spite of residual errors on each vowel (the mean error is 49 Hz for F1, 125 Hz for F2 and 170 Hz for F3), the vocalic space obtained covers well our speaker's vocalic space, in the $F_2 - F_3$ plane as well as in the $F_1 - F_2$ plane.

The essential goal is reached, i.e. the vocalic space of our subject is covered by the adapted model. It would be possible to slightly improve the fitting between natural and synthesized vowels by modifying α and β coefficients. We prefer not to modify them because this could modify the model behavior.

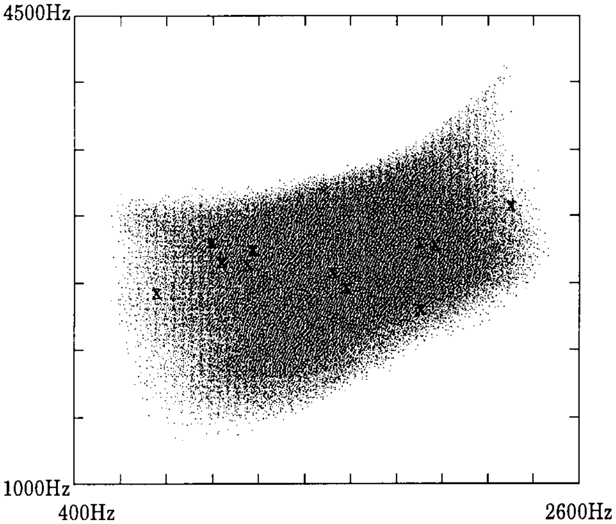

Figure 3: F2/F3 space (x points denote vowels uttered by the speaker)

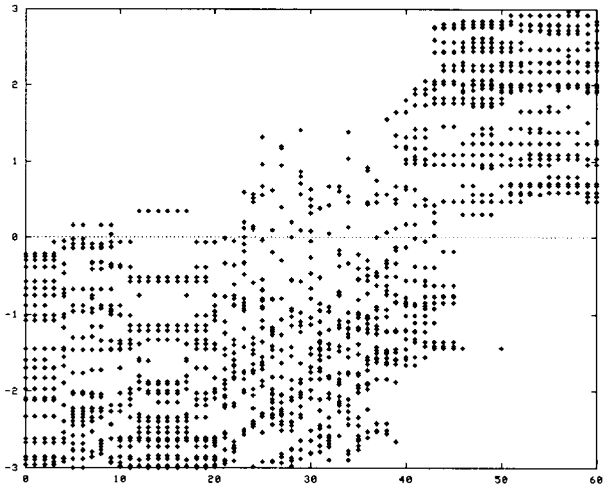

Figure 4: Articulatory trajectories for the tongue location with the codebook obtained by random sampling of the articulatory space when /iu/ is uttered. −3 (resp.+3) is for front (resp. back) positions.

4 Inversion experiments

Vocal tract shapes are recovered from results of our formant tracking algorithm [7]. As a first step we do not want to impose a parametric form to the articulatory trajectories (as a sigmoid in [3, 5]) or initial and final articulatory configurations.

The inversion process works as follows:

- First, an articulatory codebook was generated [1]. We tested two construction strategies. The first codebook was obtained by random sampling of articulatory parameters in an interval of ±3 standard deviation around their mean values. It consists of 300 000 entries. The second strategy exploits root shapes which correspond to the articulatory parameters measured in MR images of the vowels uttered by our speaker. Articulatory parameters are sampled at random in the vicinity of linear transitions between any two root shapes. In the two cases we eliminate codebook entries which correspond to a constriction area smaller than $0.2cm^2$ [2]. The second strategy has the advantage of producing vocal tract shapes which are more likely from an articulatory point of view. Therefore the second codebook only consists of 65 000 entries.

- Second, considering an unknown utterance the N codebook entries which produce the closest formants to those extracted from speech are retained. The distance used is the euclidian distance between the first 3 formants frequencies.

- Third, the best articulatory trajectory (i.e. the smoothest in this experiment) among the N^T possible (where T is the number of spectra calculated for the unknown utterance) is accepted as the inversion solution.

  We use the nonlinear smoothing algorithm proposed by Ney [10] to search for the best inversion solution. This algorithm optimizes a smoothness criterion and filters out outliers. This second aspect is particularly interesting when not any codebook entry is correct for a given formant 3-tuple. This happens when sampling of the articulatory codebook is not sufficiently discriminating. Gaps in the articulatory trajectories (i.e. instant where no appropriate codebook entry has been found) are filled in by linearly interpolating contiguous vectors.

Fig. 4 visualizes the set of possible trajectories for the tongue location parameter when /iu/ is uttered. For each formant 3-tuple we keep the 15 best articulatory vectors. For the first codebook the average error on formant frequencies for the best articulatory vector is 7 Hz for F1, 7 Hz for F2 and 10 Hz for F3. For the worst vector (among the 15 best vectors) the error remains negligible (22 Hz for F1, 18 Hz for F2, 15 Hz for F3).

Fig. 5 visualizes trajectories when the second codebook is used. The scattering of trajectories is substantially smaller than for the first codebook. Nevertheless, the overall displacement of the tongue is clearly visible in the two cases.

5 Conclusion

In this paper we described how Maeda's model can be adapted to a new speaker. The use of such a model ensures a realistic behavior during the inversion process. In spite of remaining errors on the first three formant frequencies

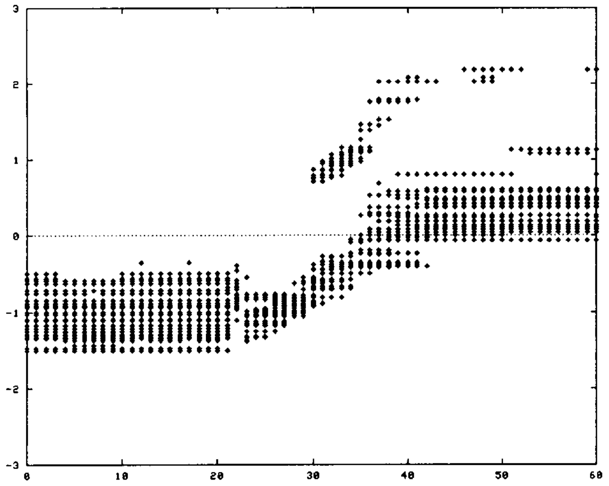

Figure 5: Possible articulatory trajectories with the codebook derived from root shapes.

the vocalic space of our speaker is well covered by the model.

We did not impose strong constraints (like the initial and final configurations of the articulators, or a parametric form) on articulatory trajectories. The first inversion results show that on the one hand the compensation phenomena are preserved (spreading of *tongue location* parameter in Fig. 4), and on the other hand the possible trajectories are consistent (/i/ is always a front vowel, whereas /u/ is a back vowel). Ney's algorithm allows us to impose a regularity constraint to choose the *best* trajectory. Nevertheless we are now working on a method to iteratively optimize the trajectories so as to alleviate problems caused by sampling of the the articulatory space.

6 Acknowledgments

The authors would like to thank Dr. Shinji Maeda for making available his articulatory model.

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. *Journal of Acoustical Society of America*, 63(5):1535–1555, May 1978.

[2] L.-J. Boë, P. Perrier, and G. Bailly. The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20:27–38, 1992.

[3] L. Candille. *Modèles de production et reconnaissance automatique de la parole.* Thèse de L'Université d'Avignon et des Pays du Vaucluse, Dec 1996.

[4] L. Candille and H. Méloni. Pilotage dynamique d'un modèle de production. In *Actes des 21èmes Journées d'Etude sur la Parole*, pages 103–106, Avignon, Juin 1996.

[5] M. George. *Analyse du signal de parole par modélisation de la cinématique de la fonction d'aire du conduit vocal.* Thèse de L'Université Libre de Bruxelles, Apr 1997.

[6] 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 Acousticsg*, page A44., 1965.

[7] Y. Laprie and M.-O. Berger. Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication*, 19(4):255–270, October 1996.

[8] S. Maeda. 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.

[9] B. Mathieu and Y. Laprie. Speaker normalization of the Maeda's model. In *Proceeding of International Workshop on Speech and Computer, SPECOM'96*, pages 167–170, St. Petersburg, Russia, October 1996.

[10] H. Ney. A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, 5(2):163–173, March 1983.

[11] B. H. Story, I. R. Tize, and E. A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *Journal of Acoustical Society of America*, 100(1):537–553, July 1996.

[12] C.-S. Yang and H. Kasuya. Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 623–626, Yokohama, Japan, September, 1994.