CONTROL OF A VOCAL TRACT MODEL BASED ON ARTICULATORY MEASUREMENTS AND ACOUSTIC OPTIMIZATION

L.Candille*, M. George**, A. Soquet** and H. Meloni*

* Laboratoire d'Informatique, Université d'Avignon, France
** Laboratoire de Phonétique Expérimentale, Université Libre de Bruxelles, Belgique

ABSTRACT - The control of the Maeda's articulatory model is realized using 2 different methods. One is based on articulatory measurements, the other consists in an acoustical optimization. V_1V_2 sequences are studied by both methods and some preliminary results are compared.

INTRODUCTION

Estimating vocal tract shape from speech signal interests many researchers for many years. This transformation seems to be useful to better understand production mechanism but also for its potential application in text-to speech synthesis and automatic speech recognition (Schroeter & Sondhi, 1994).

Direct measurements of vocal tract shape could help to better understand speech production mechanism and characteristics of the transformation between articulatory and acoustical representations. However, many problems remain with direct measurements. Since the end of the 80's, the only direct measurements at our disposal have been X-rays measurements. Now, new techniques such as magnetic resonance imaging (MRI) or articulography offer promising perspectives.

Because of the difficulty of direct measurements, many works concerning acoustic to articulatory mapping have been done (see for example, (Schroeter & Sondhi,1994) or (McGowan, 1994) for the state of the art). This transformation, so-called the inverse problem, is nonlinear and nonunique: more than one tract geometry might produce the same speech spectrum.

Although many acoustic to articulatory mapping methods have been proposed, studies which take into account of dynamics of speech or contextual acoustic and articulatory information are more unusual.

In this paper, we present the control of a vocal tract model based on both articulatory measurements and acoustic optimization.

In previous work (Candille and Méloni, 1995), a procedure has been proposed in order to infer control strategies of a sagittal cut model (Maeda, 1979) from direct articulatory measurements of the speaker.

Recently, a method that allows the estimation of articulatory trajectories (Soquet & George, 1996) from an area function model (Mrayati et al., 1988) has been developed.

On one hand, we have at our disposal direct measurements obtained with articulograph, on the other hand, a method of recovering of articulatory trajectories from formant transitions.

The purpose of this paper is to present preliminary results concerning a comparison of articulatory trajectories obtained by both methods on the same set of data (formant transitions of V_1V_2).

First, we will present the principles of both methods and then, we will discuss preliminary comparison of the performance of both methods.

METHODS

Schematically, speech production system is composed of an articulatory model and an acoustical model. The articulatory model allows to control the geometry of the vocal tract (the area function $a(x)$) with articulatory control parameters. The acoustical model simulates the physics of sound propagation in the vocal tract and allows to compute acoustical cues y from area function $a(x)$.

In this study, the vocal tract model used is the Maeda's model (Maeda, 1979) developed from radiographs of a human vocal tract. This model is characterized by 7 parameters relative to jaw (jw), tongue place (tp), tongue shape (ts), tongue tip (tt), larynx (lx), and 2 additional parameters to describe lips shape, one for the lip aperture (lh) and the other for the lip protrusion (lp). A model based on electrical equivalent is used as acoustical model (Badin and Fant, 1984). The acoustic cues used here are the first three formant frequencies.

1. Inferring control strategies of a sagittal cut model from direct articulatory measurements

This procedure takes place in 2 successive steps (see (Candille and Meloni, 1995) for more details):

- Static adaptation of the model

First, an adaptation of Maeda's model static characteristics is required in order to adjust the model for each speaker (Payan & Perrier, 1993).

For each speaker, the parameter which controls the average vocal tract length is adjusted in order to minimize the distance between the formant values obtained from the standard configuration of vowel [y] and those measured on the same phoneme uttered by the speaker.

For each oral vowel, an optimal configuration of the model has been defined both to minimize the distance between the acoustic parameters of the model and those of the speaker and also by adjusting the key geometrical variables of the area function, (place and area of the constriction (Xc and Ac) and labial area (Al)) according to the variations noted by (Boë et al., 1992) for each vowel. This speaker-adapted configuration is chosen from a set of vocal tract shapes that are built by limited variations based on a standard prototype for each vowel.

- Dynamic adaptation of the model

The second step consist of studying V_1V_2 articulatory transitions uttered by a speaker for the control of Maeda's model parameters.

Movements of different speaker's articulators (lips, tongue and jaw) during vocalic diphones are measured with an electromagnetic system. The recordings have been made for a speaker uttering a hundred V_1V_2 diphones each consisting of ten French oral vowels. For this recording, the receiver coils have been placed on the lower and upper lips to measure the labial movements, on the lower incisor tooth to measure jaw movement and at three points on the tongue (tongue tips, body and dorsum). A final receiver coil is placed on the upper incisor tooth as a reference point permitting possible data corrections. The recordings were made automatically at the Phonetic Institute of Aix-en-Provence with an electromagnetic system (Movetrack) (Teston & Galindo, 1990; Branderud, 1985). For each V_1V_2 sequence, it is possible to visualise the trajectory of the articulators in the X-Y plane.

The receiver coils movements measuring the natural articulatory activity are not directly linked to the control parameters of Maeda's model. The receiver coils movements are projected onto the X and Y axis. The behavior of those projections is monotonous but non-linear. The movements of the projections are therefore fitted with sigmoids. These functions are then used to model the dynamic of Maeda's model parameters. In the projection onto the X axis, the articulators movements are related to the model parameters in the following way:

the upper lip receiver coil movement with the protrusion parameter (lp),

the dorsum movement with the tongue position parameter (tp).

In the projection onto the Y axis, the articulators movements are related to the model parameters in the following way:

the lower lip movement with the opening labial parameter (lh),

the lower incisor movement with the jaw parameter (jw),

the apex movement with the tongue tip parameter (tt),

the dorsum movement with the tongue shape parameter (ts)

Thus the trajectory of each articulatory parameters is modelled with a sigmoid as described by equation (1):

$$u_i(t) = S_{i1} + \frac{(S_{i2} - S_{i1})}{1 + e^{-p_{i0}(t - t_{i0})}} \tag{1}$$

This model expresses the evolution in time of the ith articulatory control parameter $u_i(t)$ in terms of 2 steady states and a transition between these 2 steady states. Steady states are respectively characterized by both asymptotic values of the sigmoid S_{i1} et S_{i2}, and the transition by the slope p_{i0} and the inflection point t_{i0}. Figure 1 illustrates this model.

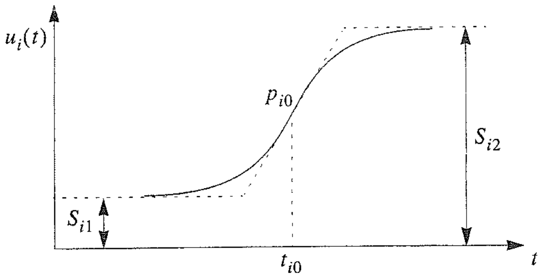

Figure 1 : Modelling V_1V_2 sequences with a sigmoids characterized by 4 parameters: steady states S_{i1} et S_{i2} ,the slope p_{i0} and the inflexion position t_{i0} .

For each V_1V_2 sequence, S_{i1} and S_{i2} values map respectively to the initial and final values of the trajectory, p_{i0} and t_{i0} values are optimized in order to minimize the distance between the natural data values and the function values. Most of the natural articulatory trajectories are modelled fairly accurately by these functions.

2. Estimation of articulatory trajectories from formant transitions

The method is similar to the one used in (Soquet & George, 1996) to obtain articulatory trajectories from an area function model (Mrayati et al., 1988).

Movement of each articulatory control parameters is modelled by a sigmoid. Indeed, this simple model is able to describe observed trajectories. Sigmoidal model used in V_1V_2 study is described by equation (1) and figure1. In our case, a sigmoid is associated to each of the evolution in time of the 7 articulatory control parameters of Maeda's model.

The analysis consists in determining parameters which govern the sigmoids. 4*7 parameters have to be determined by the analysis. The analysis requires 2 steps:

- Estimation of steady states of sigmoids(S_{i1} and S_{i2})

For each V_1V_2 sequence, S_{i1} and S_{i2} are chosen to be equal to the initial and final values of the trajectory obtained by the first method described above.

- Estimation of slopes and inflection points(p_{i0} and t_{i0})

After estimating steady states, slopes and inflection points are obtained with an optimization procedure. The principle of that optimization is presented in figure 2.

Slopes and inflection points are estimated by analysis-synthesis in order to minimize, on the whole transition, the error between target formants and those which are computed. These parameters allow to obtain evolution in time of articulatory control parameters $u_i(t)$ and of computed formant $y(t)$. These formant values are compared to target formants $y^d(t)$. An error is computed and minimized by a steepest descent algorithm by adjusting slope and inflection point values. The determination of initial conditions (initial slopes and inflection points) for the optimization is resolved by presenting different sets of slopes and inflection points to the system. For each set, error between target formants and those which are computed is estimated. The set of slopes and inflection points that minimize this error is adopted as initial conditions for the optimization.

This method has been able to analyze V_1V_2 logatomes (where V_1, V_2 are French oral vowels). The results have shown a good adequacy between measured formants and those obtained with the articulatory trajectories modelled. At the articulatory level, V_1V_2 transitions have been realistic and interpretable (Soquet and George, 1996).

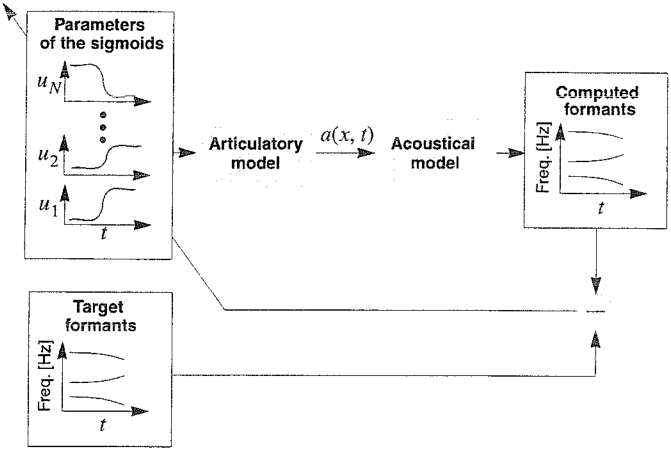

Figure 2 : Principle of the estimation of slopes and inflexion points by analysis-synthesis.

RESULTS AND DISCUSSION

Ten repetitions of a corpus of V_1V_2 logatomes (where V_1, V_2 are French oral vowels) have been recorded for one female subject. The first method has been applied to the whole corpus and results can be found in (Candille and Meloni, 1995). The second method has been applied to one repetition of the corpus.

Results of both methods on two logatomes are presented on figure 3 and 4. Those two samples are representative of the behavior of both methods.

Figure 3 presents the evolution in time of 6 Maeda's parameters (jw, tp, ts, tt, lh, lp) obtained by both methods for [a i] and [i u] transitions. Articulatory trajectories obtained from articulatory measurements and by optimization are respectively represented in plain and in narrow dashed lines.

Figure 4 presents the measured formant transitions (dashed lines), formant transitions computed by the acoustical model (Badin and Fant, 1984) from articulatory trajectories obtained from articulatory measurements (plain lines), and formant transitions resulting from optimization (narrow dashed lines).

At the articulatory level, those figures show that the optimization procedure provides stepper slopes than the articulatory derived trajectories. This observation can be explained by the values chosen for the steady states of the sigmoids (S_{i1} and S_{i2}) in the case of the optimization procedure.

Tongue place and tongue tip tend to be very similar for both methods. The slope and inflection point of other parameters appear to be very different.

At the acoustic level, the match between the target formants and those obtained with the vocal tract modelling is better with the optimization procedure than with the articulatory derived parameters. This is not surprising given the fact that the optimization is made on formant values. However, this has no implication on the likelihood of the quality of the trajectories obtained by both methods.

In order to get more insight on both procedures, it could be of interest: (i) to start the optimization from the sigmoid parameters derived from articulatory data, (ii) to ensure that the hypothesis of the optimization procedure concerning the steady states are satisfied, (iii) to study the sensitivities of the sigmoid parameters in order to better understand their influence on the formant values for particular V_1V_2 logatomes.

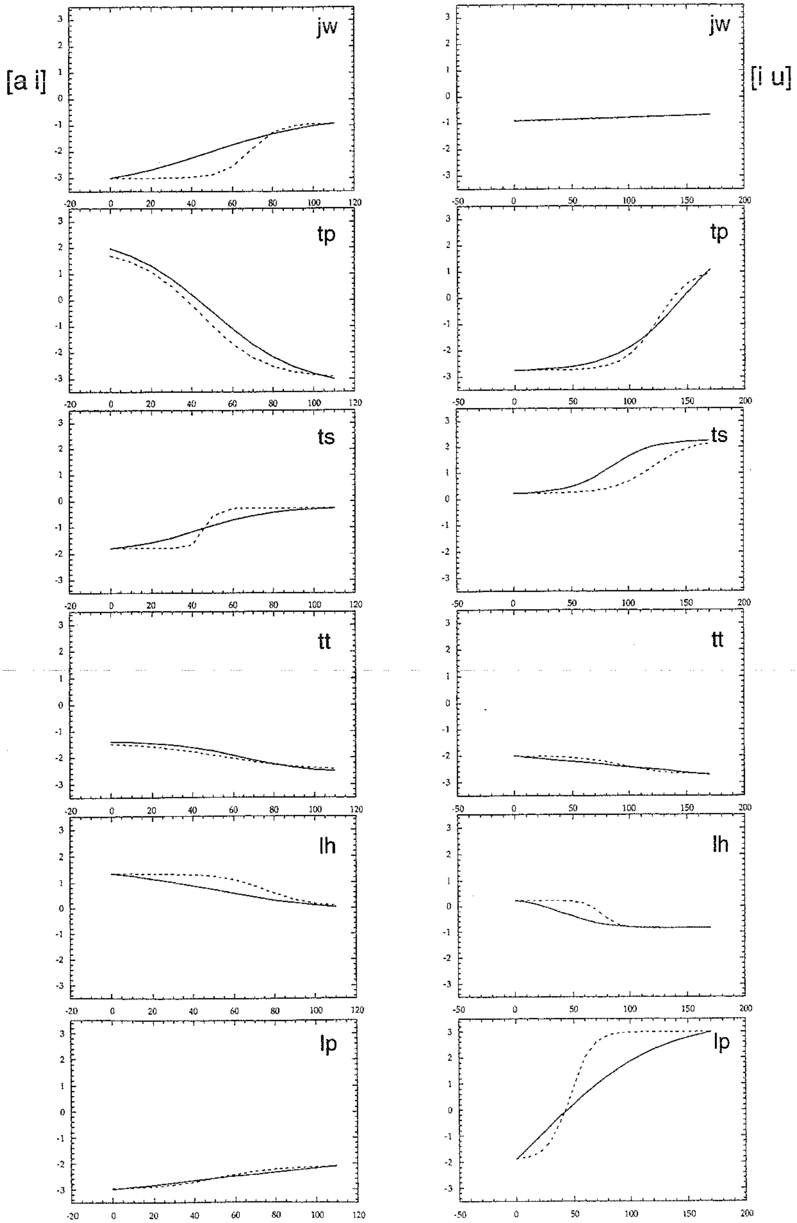

Figure 3 : Articulatory control parameter trajectories obtained by direct measurements (plain lines) and by optimization (narrow dashed lines) for [a i] and [i u] transitions.

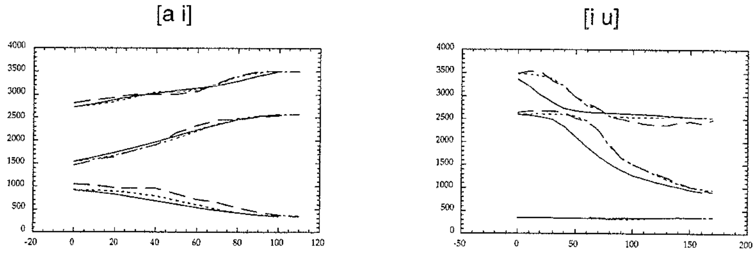

Figure 4 : Measured [a i] and [i u] formant transitions (dashed lines), formant transitions computed by the acoustical model (Badin and Fant, 1984) from articulatory trajectories obtained from articulatory measurements (plain lines), and formant transitions resulting from optimization (narrow dashed lines).

REFERENCES

Badin, P. and Fant, G. (1984) *Vocal tract frequency domain calculation technique*, R. Inst. Tech. Speech Trans. Lab., Q. Prog. Stat. Rep. 2-3, 53-108.

Branderud, P. (1985) *Movetrack, a movement tracking system*, in Proceedings of the French-Swedish Symposium on Speech, GALF, Grenoble, France, 113-122.

Boé, L.J. Perrier, P. and Bailly, G. (1992) *The Geometric Vocal Tract Variables Controlled for Vowel Production: Proposals for Constraining Acoustic-to-Articulatory Inversion.* Journal of Phonetics 20, 27-38.

Candille, L. and Meloni, H. (1995) *Automatic speech recognition using production models*, ICPhS' 95 Stockholm, vol. 4, 256-259.

Maeda, S. (1979) *Un modèle articulatoire de la langue avec des composantes linéaires*, Actes des X^{es} Journées d'Etude sur la Parole, 154-162.

McGowan, R.S. (1994) *Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests*, Speech Communication, vol. 14, 19-48.

Mrayati, M. Carré, R. and Guérin, B. (1988) *Distinctive regions and modes: a new theory of speech production*, Speech Communication, vol. 7, 257-286.

Payan, Y. and Perrier, P. (1993) *Vowel normalization by articulatory normalization: first attempts for vowel transitions*, Eurospeech 93, vol.1, 417-420.

Schroeter, J. M. and Sondhi, M. (1994) *Techniques for estimating vocal-tract shapes from the speech signal*, IEEE Transactions on Speech and Audio Processing, vol. 2, n°1, part II,133-150.

Soquet, A. and George, M. (1996) *Estimation de trajectoires articulatoires à partir de transitions formantiques: Application à l'analyse de séquences V_1V_2 et V_1CV_2*, Actes des XXI^{es} Journées d'Etude sur la Parole, 91-94.

Teston, B. and Galindo, B. (1990) *Une station de travail d'analyse de la production de la parole*, Actes des XVIII^{es} Journées d'Etude sur la Parole, 180-184.