DYNAMIC CONTROL OF A PRODUCTION MODEL

Laurence CANDILLE, Henri MELONI

Laboratoire d'informatique d'Avignon 339, chemin des Meinajariès BP 1228
84140 AVIGNON Cedex FRANCE
tél.: (33) 90 84 35 09, fax: (33) 90 84 35 01, e-mail: candille@univ-avignon.fr

ABSTRACT

A number of experiments have shown that it is possible to use production models for speech recognition tasks [6] and [2]. We present here the first results of an adaptation of Maeda's statistic model. We have also demonstrated the importance of taking into account the static and dynamic characteristics of the speaker. Some preliminary results for the identification of V1-V2 sequences are also provided.

1. INTRODUCTION

In the context of automatic speech recognition based on speech production models, this paper presents the first known results of an adaptation of Maeda's statistic model [4]. We also demonstrate the importance of taking into account the static and dynamic characteristics of the speaker. For each speaker, the parameter related to the total vocal tract length is adjusted. For each oral vowel, an optimal configuration of the model is defined. The production of the vowel-vowel formantic transition by a linear interpolation of the parameters between both target configurations does not always permit the precise reproduction of the speaker's trajectory. Consequently, we have made measurements of the different natural movements of the speaker's articulators (lips, tongue and jaw) during vocalic diphone utterance. The transposition into the model of the velocities and the accelerations measured at various articulator points produces acoustic transitions that are very close to those of the speaker. Additionally, contextual strategies deduced from natural data examination are applied to the dynamic control of the model parameters to recognizing vocalic diphones.

2. STATIC ADAPTATION OF THE MODEL

Maeda's articulatory model is a statistical model developed from radiographies of a human vocal tract. An adaptation of its static characteristics is required in order to adjust the model for each speaker [5]. This is described in the following sections.

2.1 Modification of the vocal tract length

For each speaker, the parameter related to the total vocal tract length is adjusted so that the acoustic space produced by the model matches that of the speaker as closely as possible. This adjustment is made by minimizing the distance between the formant values obtained from the standard configuration of vowel [y] and the values calculated for the utterance of this phoneme by the speaker. This distance is defined as being the number of critical bands between the model's value and that of the speaker (for each formant).

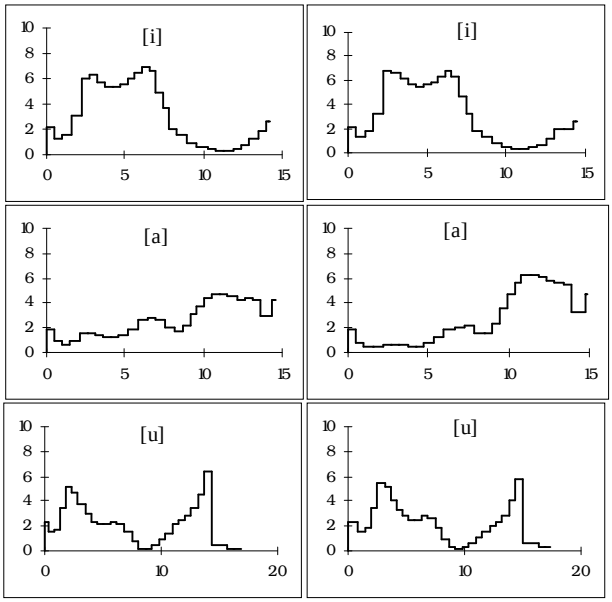

Figure 1: Comparison of the area function obtained with length adapted Maeda's model (right) and those corresponding to the standard configurations (left) for the three cardinal vowels [i], [a] and [u]. The X-axis represents the distance from the glottis (cm), the Y_axis represents the regions' area (cm^2).

2.2 Choice of optimal configuration

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 [1] for each vowel (figure 1). 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. The quality of the vowels obtained is also validated by ensuring the synthetic quality of the sounds produced.

In other respects, the configurations have been chosen in such a way that control variations of the model, when passing from one configuration to another, are as close as possible to the natural movements of the speaker's articulators when producing the same sequence.

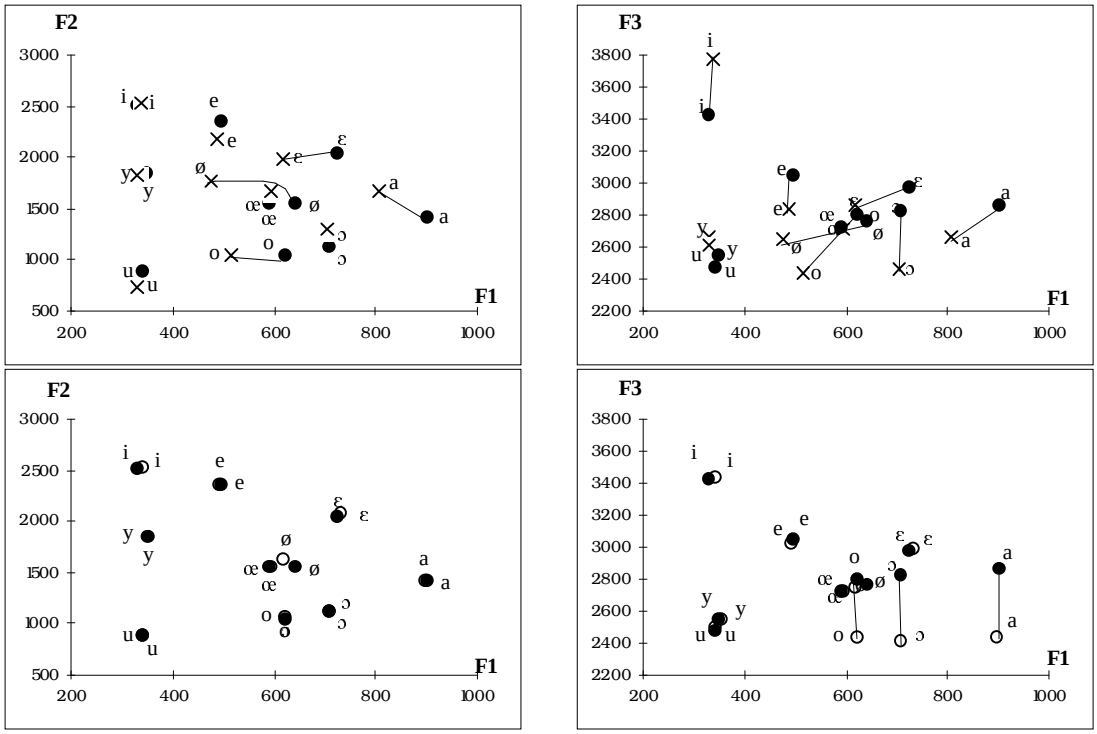

Figure 2: Comparison in the F1/F2 (left) and the F1/F3 (right) planes of the French oral vowel acoustic values; the speaker's values are represented by character "•", those of the model standard vowel by "+" (upper figures) and those corresponding to adapted configurations by "o" (lower figures).

Figure 2 presents a comparison in the F1/F2 and the F1/F3 planes of the acoustic values of the speakers mean vowel formant values with those obtained by the model (both with respect to the standard configurations and with respect to the adapted configurations). The model length coefficient was first fixed for each configuration and the ten standard prototypes used are a subset taken from the 33 vowels of UPSID [3] and [8]. In the F1/F2 plane, the French oral vowels of the two speakers tested are well represented by the formantic values of the adapted configurations. In the F2/F3 plane, we note, for both speakers, the third formant value of the back vowels [a], [ɒ] et [ɔ] is not reached by the acoustic values of the model using adapted configurations. Our results show that the sum of the acoustic distances between the vowels of the model's optimized configurations and the speaker's vowels is improved by more than 70% compared to the same sum using the standard vowels.

3. DYNAMIC ADAPTATION OF THE MODEL

At the present time, the model includes the static representation of all French vowels but does not include a method to go from one configuration to another. The production of the vowel-vowel formantic transition by a linear interpolation of the parameters between both target configurations does not, in some cases, reproduce precisely the trajectory obtained for a speaker (figure 4).

We have also studied the acoustic effects (in the F1/F2 plane) produced by a variation of the velocity of each articulator. The results are clearly improved when the articulator movements are as close as possible to reality. The transition from one vowel to another depends on the coordination and the velocity of the model control parameters. We propose to adapt the model to the articulatory strategies used by a speaker by making the hypothesis that these transitions have an optimal acoustic behaviour. We have made measurements of the different natural movements of the articulators (lips, tongue and jaw) during vocalic diphone utterance.

3.1 Articulator movement recordings

The recordings have been made for a speaker uttering a hundred V1-V2 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) [7]. For each V1-V2 sequence, it is possible to visualise the trajectory of the articulators in the X-Y plane.

3.2 "Natural" control of the model

The receiver coils movements measuring the natural articulator 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. 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)

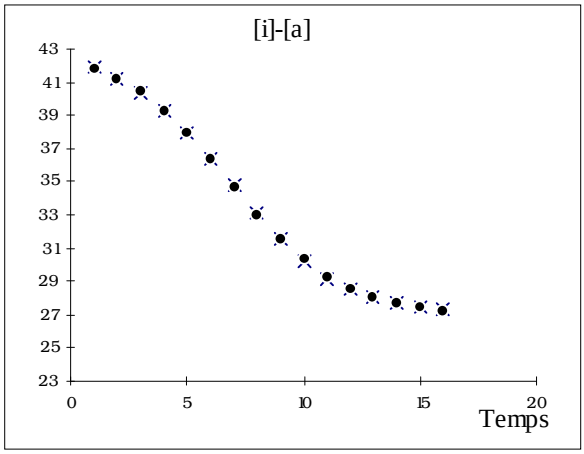

Figure3: Comparison between "tongue shape" natural articulator trajectory ("×") and the optimised sigmoïdal curve modeling this movement ("•"). Speaker "lc" uttered [i]-[a] sequence.

Taking [i]-[u] as an example: the model transition in which the parameter values have been linearly interpolated was found to provide a poor match with the speaker's transition (fig 4). For this sequence and this speaker we note that jaw movement is unimportant. When the articulator's real movements are not sufficiently important, the interpolation of the corresponding parameter is linearly interpolated if necessary.

For each V1-V2 transition, we would like to model the natural movement of each articulator obtained from articulatory recordings.The behaviour of these trajectories is monotonic but not-linear. We propose to model these trajectories with sigmoïds, these functions are described by equation (1):

$$U\ (t) = S1 + \frac{(S2 - S1)}{1 + e^{-b(t - t_0)}} \qquad (1)$$

This model represents the articulator's behaviour as a function of time. Variables S1 and S2 represent the asymptotic values of the two stable parts of the curve, b and t0 values are respectively the slope and the bending point of the curve.

For each V1-V2 sequence, S1 and S2 values map respectively to the initial and final values of the trajectory, The b and t0 values are optimised 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. Figure 3 represents the comparison between the "tongue shape" articulator trajectory in time when the passage from vowel [i] to vowel [a].

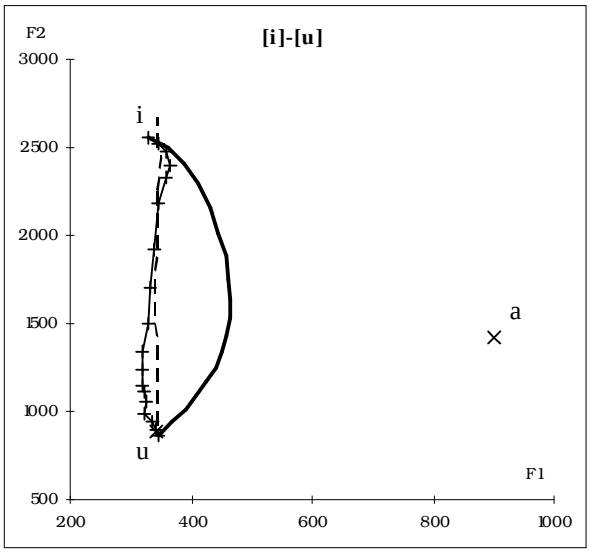

Figure 4: Comparison of the acoustic effects in the F1/F2 plane during the transition from configuration [i] to configuration [u] with "linear" control (thick black line) and "natural" control (character "+"). The curve plotted with character "-" represents the real trajectory of speaker lc formants.

Figure 4 compares the acoustic effect in the F1/F2 plane when we move from configuration [i] to configuration [u] with a linear interpolation of the model parameters ("linear" control) and also with an interpolation of the model parameters deduced from the articulator's movement recordings ("natural" control). The trajectory characterizing the speaker's production is the ideal reference which is to be matched. We carried out the same experiment with all V1-V2 vocalic transitions. For each sequence, we first identified the active articulators during natural vocalic transition, the other ones are linearly interpolated between both model corresponding configurations.

The acoustic results obtained in the F1/F2 plane are better than the formantic trajectories resulting from a linearly interpolated model for the [i]-[u], [e]-[u], [ɛ]-[y], [y]-[ɛ], [y]-[a] et [y]-[u] sequences including a remarkable improvement for [i]-[u] (fig. 4) (formant crossing occurs). On the other hand, the correlation has

been reduced (compared with that of the linear control base case) for the [i]-[o], [e]-[ɔ], [ɛ]-[o], [ø]-[ɔ], [o]-[y], [ɔ]-[i] and [u]-[y] transitions. For all other sequences results are virtually the same. A number of the diphones obtained with the natural control of the model parameters have been synthesized and aurally validated.

3.3 Application to vocalic diphones recognition

We present in this part, the first results of the oral vowel sequences identification by using contextual strategies deduced from a study of the speaker's natural articulator movements in order to control the model.

Articulator measurements were made on a female speaker when she uttered a hundred of V1-V2 sequences each consisting of ten French oral vowels. The acoustic and physiological signals were measured simultaneously. The model inversion process used and our recognition strategy have both been described in [2]. In order to optimize the static search for configurations which constitutes the first part of our recognition strategy, we have generated a codebook i.e a table of acoustic vectors and corresponding articulatory vectors providing initial and final configurations for each transition. The method used was the same as described in [2], three different articulatory tables are used here. Table T0 contains the ten standard configurations of the model. Table T1 contains the ten optimized configurations (§ 2.2). Table 2 contains 20 configurations per vowel; each one representing a different vowel in context.

Having both the articulatory tables and a set of sigmoïd functions to model the trajectory of each model control parameter corresponding to each V1-V2 sequence, we can now start our recognition strategy. The recognition rate in first position of the V1-V2 transitions, for speaker lc, and for each working-table is stated in Table 1.

	T0	T1	T2
recognition rate	50%	70%	61%

Table1: The recognition rate in first position of the V1-V2 transitions, for speaker lc (across), and for each working-table (down). The three tables T0, T1, T2 contain respectively the 10 standard configurations, the 10 optimized configurations and 20 contextual configurations.

Results

These first experiments testing the use of articulatory measures for the control of a production model show that articulator movements can be modelized by sigmoïdal functions and the use of these models improves (in some cases) the acoustic results when passing from one vocal tract shape to another. Some results are very interesting (fig. 4) and some improvements can be made. Indeed, the recognition rate actually obtained are very interesting. Our study shows clearly that speaker adaptation is necessary and significantly improves the results.

During our experiments, we used only articulator movements. We would like to use also the relative position of the articulators. Consequently, we propose to improve our technique 1) by modifying our speaker adaptation strategy, to find an optimal configuration per vowel. An optimisation of the whole model parameter set must be carried on, in order to optimize the correspondence between the speaker's and the model's vocal tracts. 2) concerning the model's dynamic control, the modelling of the transition from one configuration to another must not be applied to the model control parameter, but rather to the direct control of the sagittal function variables corresponding to the receiver coils position on the speaker tongue.

4. CONCLUSION

The Maeda's model study leads us to specify the capacities of this production system within the framework of vocal recognition. A model adaptation to the static and dynamic speaker characteristics is necessary and significantly improves the results. In fact the adaptation to several speakers (both male and female) allows us to obtain reference vowels in which the acoustic distance sum of the speaker phonemes is improved by more than 70%. The input into the model of the velocities and the accelerations measured in various articulator points allows us to produce acoustic transitions that are very close to the speaker's. We applied these contextual strategies to the dynamic control of the model parameters in a vocalic diphone recognition process. The results we obtained are very interesting but could be improved upon. Our speaker automatic adaptation strategy, must be modified. The contextual strategy we obtained must be directly applied to the control of the sagittal function variables. We envisage making further tests on several speaker and we also intend to use this technique for the identification of some consonants.

5. REFERENCES

1. Boë, L.J., Perrier P. et 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.
2. Candille L. and Méloni H. (1995) "Automatic speech recognition using production models" ICPhS' 95 Stockholm, vol. 4, 256-259.
3. Maddieson, I. (1986), *Patterns of Sounds*. 2nd Edition, Cambridge University Press.
4. Maeda S. (1979), "Un modèle articulatoire de la langue avec des composantes linéaires", Actes des 10es *JEP*, 154-162.
5. Payan Y. et Perrier P., (1993), "Vowel normalisation by articulatory normalisation: first attemps for vowel transitions", *Eurospeech 93*, vol.1, 417-420.
6. Rose R.C., Schroeder J. and Sondhi M.M. (1994), "An investigation of the potential role of speech production models in automatic speech recognition". In Proc. Int. Conf. Sp. Lang. Proc., vol. 2, 575-578.
7. Teston B., Galindo B. (1990), "Une station de travail d'analyse de la production de la parole", 18es *JEP de la SFA*, Montréal, 28-30 Mai 1990, 180-184.
8. Vallée N. (1994) *Systèmes vocaliques: de la typologie aux prédictions*, Thèse Doc. ICP Grenoble.