ADAPTATION OF NATURAL ARTICULATORY MOVEMENTS TO THE CONTROL OF THE COMMAND PARAMETERS OF A PRODUCTION MODEL

Laurence CANDILLE, Henri MÉLONI
Laboratoire d'Informatique, Centre d'Enseignement et de Recherche en Informatique 339 chemin des Meinajariés, BP 1228, 84911 AVIGNON Cedex 9 FRANCE Tel (33) 4 90 84 35 09 laurence.candille@univ-avignon.fr

ABSTRACT

A number of experiments have shown the importance of the use of speech production models for automatic speech recognition ([1],[4],[6]). This work is very interesting for the concise representation of the sound coarticulation phenomena in continuous speech. Maeda's statistical model [5] has been chosen to conduct our experiments. The first part of the paper focusses on adjusting the model configurations characterizing the French vocalic sounds in an optimum way so as to minimize the acoustic distances from the phonemes produced by a speaker. The second part provides a control strategy for Maeda's model command parameters.

1. INTRODUCTION

To build up reference configurations optimally adapted to the speaker, several techniques have been used : modifications of the length of models, shifting of the symmetry axis, geometrical transformation of Maeda's model according to the structure of the vocal tract of the speaker (obtained by radiographies). The modifications brought to the static component of the model have allowed to obtain very close acoustic spaces for the model's productions and the speaker's utterances.

Maeda's model does not provide a method to go from one vocal tract shape to another. A process enabling to control its pseudo-articulatory movements has to be added to the basic model. For this, we have made some articulatory trajectory measurements for one speaker uttering vocalic logatoms and some short sentences. These data recordings have been made with an electromagnetic system (Movetrack) which allows to follow the movements of the receiver coils placed on the articulators (lips, tongue, jaw). The natural trajectories are represented by a set of sigmoïdal functions used for the dynamic control of the model. The advantage of this representation is that it accurately follows the evolution of the movements of the speaker's articulators and that it quantifies their main characteristics correctly. Furthermore, this study yields interesting results on articulators desynchronization during the utterance of some phonemic sequences.

2. MODEL'S ADAPTATION TO THE STATIC CHARACTERISTICS OF THE SPEAKER

Articulatory models are generally used to produce synthetic speech and, in this context, it is essential to obtain sounds which are easy to identify and pleasant to hear. To take these constraints into account, several authors propose reference configurations that adjust the model's parameters to represent vowels [8]. In many cases the acoustic characteristics of the "standard" phonemes are very far from the same sounds uttered by different speakers. One of our main tasks has consisted in adjusting the characteristics and parameters of Maeda's model so as to propose configurations that minimize the acoustic distance between the sounds uttered by a speaker and those produced by the model. The distance used gives a perceptually equivalent weight to each of the first three formants.

2.1. Optimal adjustment of the model's length

Maeda's statistical model has been developed from radiographies of a human vocal tract. The examination of several speakers' radiographies has led us to develop three adaptation methods :

- global and optimal modification of the vocal tract's length,
- separate modifications of front and back vocal tract's parts,
- optimal adjustment by superimposing the model's and the speaker's vocal tracts.

After a series of tests performed with the whole vowel set, vowel [y] has been chosen to optimally adapt the model's total length. Results of this first method show that the sum of acoustic distances between the vowels of the standard configurations adapted in length and the speaker's vowels is reduced by 70% compared to the same sum using the standard non-adapted vowels.

Separate modifications of the front and back vocal tract's parts of the model based on acoustic observations for vowel [y] have allowed to improve previous results (reduction of 78% of initial measures).

Radiographies of a female speaker's vocal tract have allowed us to transform Maeda's model's vocal tract in such a way that the speaker's and model's sagittal cuts match as precisely as possible. In this case, the improvement rate is similar to the previous one (77%).

2.2. Optimization of standard oral vowels configurations

After adapting the model's length to the speaker, an optimal configuration is defined for each oral vowel in order to minimize the distance between the acoustic productions of the model and those of the speaker. The optimization process has been described in [3].

Following this optimization operation, the distance between the model's and the speaker's acoustic spaces (the sum of acoustic distances between optimized vowels configuration of the model and vowels uttered by the speaker) is again significantly reduced : 83% of reduction obtained with the first length adaptation method, 85% of reduction obtained with separate modifications of the front and back vocal tract's parts of the model and 80% of reduction with the geometrical transformation of the model's vocal tract.

In the F1/F2 plane, the French oral vowels of a speaker are acoustically well represented by the optimized configurations of the model which is adapted by separate modifications of the front and back parts of the vocal tract. However, we can note that in the F1/F3 plane the third formant value of back vowels [a], [o] and [ɔ] is not accurately reached . These results could be improved with an optimal adjustment to the speaker of the transfer function from the sagittal cut to the area function. This operation would be difficult to implement in the scope of automatic speech recognition. The method which consists in adapting separately the front and back parts of the vocal tract is both simple and efficient ; it will be used in the following parts.

3. STUDY OF THE ARTICULATORY MOVEMENTS OF A SPEAKER

At the present time, Maeda's model includes the static representation of all French vowels but it does not include a method to go from one configuration to another. To solve this problem, the first test has consisted in linearly interpolating the command parameters of the model between both target configurations. In many cases, the formantic trajectories observed for a speaker are not accurately reproduced by the model so controlled (fig. 4).

When the distance between the trajectories produced by the linearly interpolated model and those observed for the speaker is too large, we propose to transfer the articulatory strategies used by the speaker into the model [3]. For this, we have made measurements of the different natural movements of the articulators (lips, tongue, jaw) with an electromagnetic system : Movetrack ([2], [3], [7]). Observations are made from a corpus composed of sequences of all French oral vowels uttered by a speaker. For the study of the temporal behaviour of the receiver coils, the movements are projected onto axes X and Y in the following way :

- the upper lip movement characterizes the protrusion on the horizontal axis,
- the lower lip movement characterizes the labial distance on the vertical axis,
- the jaw movement is projected onto the vertical axis,
- the movement of the tongue's middle coil is projected onto the vertical axis to characterize the opening at the occlusion point,
- the movement of the tongue's back coil is projected onto the horizontal axis to characterize the place of the occlusion point.

In this representation, the receiver coils trajectories can easily be modelled by logistic functions (sigmoïdal) which characterize the movements velocity and acceleration. After normalizing the parameters values in relation to the extreme positions of the receiver coils, the examination of the trajectories shows significant differences with regard to the slope and the bending point position of the curves (fig. 1).

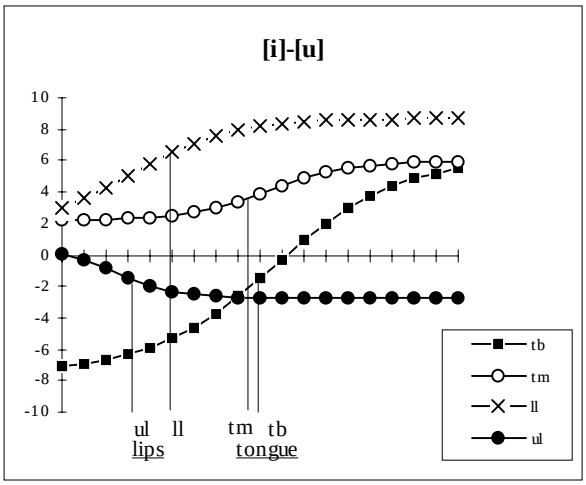

Fig. 1. Comparison in time of the bending points (vertical line) of the sigmoïdal functions which represent the receiver coils movements associated to the lips and to the tongue for speaker LC uttering sequence [i]-[u]. Symbol « × » represents the labial height, symbol « ● » represents the protrusion, symbol « ▪ » represents the movement of the tongue's back and symbol « ○ » represents the movement of the tongue's middle.

The bending points of the curves characterize the halfway time of the trajectories. For transition [i]-[u] (fig. 1), the comparison of these variables shows clearly that the labial movements (labial height and protrusion) start earlier than the tongue receiver coils movements (back and middle). This phenomenon appears systematically for all transitions from non-rounded front vowels to rounded back vowels. On the other hand, in symmetrical situations such a regular behaviour is not observed. In cases when the articulators are significantly activated and the maximal velocity point of the movement is shifted in relation to the median temporal place of the transition, the diphons for which this phenomenon is clearly marked (black squares) are differentiated from those for which it is less marked in figure 2 (grey squares). White squares on figure 2 represent two situations :

- diphons for which the articulators movements are not marked,
- diphons for which the articulators movements are marked but without any desynchronization (ex. : [e]-[ɔ], [ɛ]-[ø], [ɛ]-[ɔ], [o]-[y]).

4. CONTROL OF MAEDA'S MODEL'S PARAMETERS

4.1. Links between articulatory data and the model's parameters

So as to control Maeda's model's parameters in a both more natural and particularly more acoustically efficient way, we have introduced the desynchronization phenomenon observed for a speaker in the command process of these parameters. This new command – which we will call « natural command » in contrast with the « linear command » – is implemented by associating the receiver coils used to the model's parameters in the following way :

- the upper lip horizontal movement with the lip protrusion parameter *(lp)*,
- the lower lip vertical movement with the labial height parameter *(lh)*,
- the lower incisor movement with the jaw parameter *(jw)*,
- the vertical movement of the tongue's middle receiver coil with the tongue shape parameter *(ts)*,
- the horizontal movement of the tongue's back receiver coil with the tongue position parameter *(tp)*.

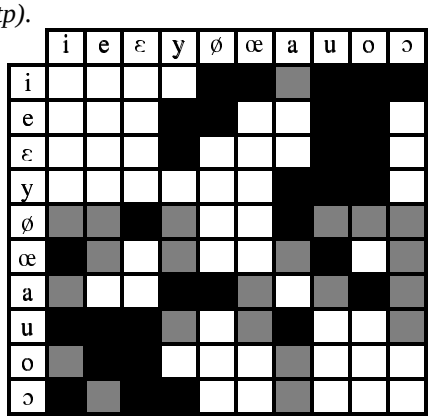

Fig. 2. Division of the diphons (row-column) uttered by a speaker (LC) into three classes which characterize the articulators behaviour corresponding to the observed transitions. Black squares correspond to transitions whose articulators movements present important desynchronizations in relation to a linear movement. Grey squares correspond to transitions whose articulators movements have small desynchronizations in relation to a linear movement. White squares correspond to transitions whose articulators movements look like a linear movement.

The receiver coil measuring the tongue tip activity is mainly useful to model phonemic sequences including consonants. It has not been used in this study limited to vocalic diphons. The corresponding parameter *(tt)*, as well as the parameter related to the larynx *(lx)*, are linearly interpolated.

4.2. Transitions production and results

For a V1-V2 transitions production with Maeda's model, we look for the optimal configurations corresponding to the target vowels and we apply both commands (natural and linear) to move the parameters from one configuration to another. Six series of transitions each consisting of one hundred diphons uttered by one female speaker have been used for the experiments. For more than half of the diphons, the linear command allows to follow quite precisely the trajectories of the speaker's utterances in the acoustic space (fig. 3). In these situations it is not necessary to use a more sophisticated method. On the other hand, for the remaining transitions, the natural command allows to reduce significantly the distance to the real acoustic trajectories in half of the cases (fig. 3). These contexts correspond, among others, to important articulatory movements (diphons composed of non-rounded front vowels and back vowels) involving several articulators. For the remaining quarter of diphons, none of these commands (equivalent in this case) allow satisfactory modelling.

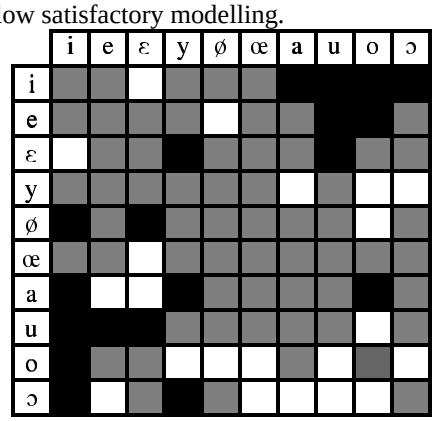

Fig. 3. Division of the diphons into three classes which characterize the observed transitions behaviour in relation to the ones produced by both command types of the model. Grey squares correspond to transitions accurately modelled with a linear command. Black squares correspond to modelled transitions whose distance to the real trajectories is significantly reduced by the natural command. White squares correspond to the diphons for which both commands have a similar behaviour and the obtained results do not accurately match the real transitions.

These results have been obtained in an automatic way, notably for the formant values calculation, and this process, in some situations, can produce erroneous measures. A bad allocation of phonetic targets configurations can also induce big acoustic distances, but the general appearance of the formantic trajectories is fairly good. Improvements obtained with the natural command are sometimes very spectacular as in the case of transition [i]-[u] where the value of the first formant is constant during the whole movement (fig. 4) and where the crossing between formants F2 and F3 (characteristic of this diphon) is correctly represented by the model.

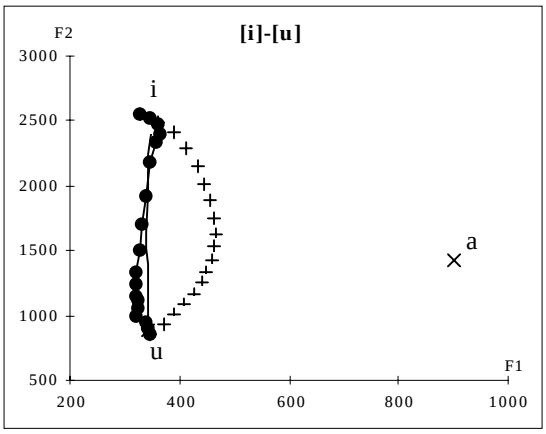

Fig. 4. Comparison of the acoustic effect in the F1/F2 plane during the transition from configuration [i] to configuration [u] with « linear » control (character « + ») and « natural » control (character « ● »). The continuous line represents the real formantic trajectory for speaker LC.

5. CONCLUSION

This study emphasizes the interest of the use of the production model for the representation of sound coarticulation phenomena. It is first necessary to adapt the model to the static characteristics of the speaker so as to adequately bring both acoustic spaces closer. This operation is easy to carry out and leads to satisfactory results for the speakers of our corpus.

Some characteristics of some phonemes (formant F3 of back vowels) are still a little far from real values for several speakers.

These results could be improved with an optimal adjustment to the speaker of the transfer function from the sagittal cut to the area function, but this operation is not easy to generalize to the whole speaker set.

The study of a speaker's articulatory movements has allowed to highlight the transitions for which we observe desynchronizations of some articulators in relation to linear movements. These observations justify the Maeda's model commands implementation including these phenomena. The « natural » command of the model allows us to obtain satisfactory results for a large part of the transitions badly modelled by the « linear » command. However, a number of cases remain for which both solutions proved to be insufficient.

The model and its commands have been used for an experiment of vocalic diphons series recognition uttered by a male and a female speaker. The identification results (80% of accurate recognition) are interesting enough to justify the development of this research.

We propose to carry on with this research with a view to addressing the following questions :

- improvement of the static speaker-to-model adaptation process,
- increase in the number of speakers especially with a view to going deeper into the study of the dynamics of natural articulators by looking for regularities among the articulatory behaviours (development of a limited set of command strategies adapted to some speaker types),
- improvement of the transfer of the observations made on speakers articulators to Maeda's model's parameters,
- use of acoustic parameters easier to measure automatically than formant values,
- use of the model for recognition of phonemic sequences including consonants.

6. REFERENCES

[1] C. S. Blackburn and S. J. Young (1995), "Toward improved speech recognition using a speech production model", *Eurospeech'95*, Madrid, 1623-1626.

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

[3] L. Candille and H. Méloni (1996), "Dynamic control of a production model", *ICSLP96*, Philadelphia, 3-6 octobre, USA, vol. 4, pp. 2305-2308.

[4] L. Deng and D. Sun (1994), "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features", *J. Acoust. Soc. Am.*, 95, 2702-2719.

[5] S. Maeda (1979), "An articulatory model of the tongue based on a statistical analysis", *Journal of the Acoustical Society of America*, Vol. 65(A), S22.

[6] R. C. Rose, J. Schroeter, M. Sondhi (1996), "The potential role of speech production models in automatic speech recognition", *J. Acoust Soc. Am.* 99 (3), 1699-1709.

[7] B. Teston and B. Galindo (1990), "Une station de travail d'analyse de la production de la parole", 18es *JEP de la SFA*, Montréal, 28-30 Mai 1990, pp. 180-184.

[8] N. Vallée (1994), *Systèmes vocaliques : de la typologie aux prédictions*, Thèse de Doctorat de l'ICP, Grenoble.