AUTOMATIC SPEECH RECOGNITION USING PRODUCTION MODELS

Laurence CANDILLE, Henri MELONI

Laboratoire d'informatique
Faculté des Sciences, 33 rue L. Pasteur, 84000 avignon
tel.: 90 14 44 21
e-mail: candille@univ-avignon.fr

ABSTRACT

We present how the Distinctive Region Model (DRM) may be used for the recognition of two vowel sequences. The process studied here takes into account the characteristics of the speaker, the phonetic context and the variation of the formants during the V1-V2 transition. On average, 80% of the V1-V2 sequences are correctly identified. The article presents the results obtained for the ten French vocalic vowels.

1- INTRODUCTION

In order to verify whether speech production models are appropriate for automatic speech recognition, we primarily use the DRM model because it is simple and easy to control.

This model offers the advantage of proposing a dynamic modelisation of articulators motion to go from one vocal tract configuration to another. Moreover, it is speaker independent. The model parameters are the areas of the regions and the total length of the tube.

In this preliminary stage, we consider the identification of V1-V2 sequences uttered by several speakers. The model derived formant transition are compared with the acoustically measured ones.

2- DRM MODEL INVERSION

The aim of the present work is not to validate the model nor to modify it to solve particular difficulties. We use it as it has been designed by its conceptors [2]. The DRM model inversion process and our recognition strategy have been described in [1].

2.1 - speaker adaptation

In order for the model acoustic space to better match the speaker's, we must either modify some characteristics of the model or normalise the speaker's parameters.

The DRM model characteristics allow a speaker adaptation by varying the total tube length.

The model may be adapted in two ways: either by fixing the total vocal tract (VT) length for each speaker and is identical for every vowel or by fixing the length for each vowel of each speaker.

2.2 - codebook generation

In order to optimize the static search for configurations which constitutes the first part of our recognition strategy, we generate a codebook i.e a table of acoustic vectors and corresponding articulatory vectors which provides starting and final configurations for each transition.

The configurations in the codebook are produced by varying around a reference VT model. There is a reference vocal tract model for each vowel. The variation allowed around each reference configuration is fixed. All other configurations are produced by moving from one extreme configuration to another along a straight line in the parameter space. We also vary the interpolation type and the configuration total length.

Thus we obtain a reference table (TR -0) containing about 15 000 configurations which describe the whole vowel set.

Some VT models are associated with acoustic parameters which do not match those of the currently studied vowel, therefore they must be filtered out. By filtering the table TR-0, we create six different tables distinct from each other by the number of configurations for each vowel, the choice of these configurations and their total length. These tables will be used directly for the recognition of the V1-V2 sequences. The first working-table TT-1-1 contains the configurations of the reference table TR-0. The total length of the configurations is fixed for each speaker and is identical for each vowel. TT-1-2 contains 20 configurations per vowel (from TR-0); these configurations are speaker dependent and represent every vowel in context. Finally in table TT-1 -3 the configurations are speaker dependent and each vowel is represented by one configuration only.

The last three tables TT-2-1, TT-2-2, TT-2-3 are built respectively like the first three ones but the configurations length is fixed for every vowel of every speaker.

2.3 - recognition strategy

Firstly we measure the first three formants at the onset and the offset of the input signal. Refering to the codebook, we determine a VT configuration for each hypothesis concerning V1 and V2. All possible formant transitions are then calculated using the speaker adapted model with its two commands and different interpolation types. Secondly we identify the V1V2 sequences whose entire formant transition matches best the formant transition measured on the input signal.

3- RESULTS

3.1 - symetric model

Three male French speakers uttered a hundred of V1-V2 sequences consisting of the ten French oral vowels.

The standard DRM model configurations, even with varying lengths, yield no more than 50% recognition rate on average for the three speakers (these results must be compared with tables TT-1-3 and TT-2-3).

The recognition rate in first position of the V1-V2 transitions, for each speaker, and for each working-table is stated in Tab1.

Tab 1: The recognition rate in first position of the V1-V2 transitions, for each speaker (across), and for each working-table (down). The first three tables TT-1-1, TT-1-2, TT-1-3 contain respectively about one hundred configurations, 20 configurations and 10. The total length of the configurations is fixed for each speaker, and is identical for each vowel. The last three tables TT-2-1, TT-2-2, TT-2-3 are built respectively like the first three ones but the configuration length is fixed for every vowel of every speaker.

	TT 1-1	TT 1-2	TT 1-3	TT 2-1	TT 2-2	TT 2-3
TS	52%	67%	69%	65%	80%	80%
FB	42%	56%	55%	54%	77%	80%
PG	50%	65%	68%	50%	79%	81%
TOT	48%	63%	64%	56%	79%	80%

These first results show that the VT model length adaptation for each speaker increases the recognition rate and confirms that the standard DRM configurations are not optimal for V1-V2 sequence recognition.

As for recognition rate, tables TT-2-2 and TT-2-3 give acceptable and consistant results.

However, we note that the configurations of the table TT-2-2 allow the model to produce trajectories closer to natural speech. Table TT-2-2 contains configurations selected using formant patterns from vowels in context for each speaker, most of the time this table provides optimized initial and final configurations of the transition and therefore, the corresponding total acoustic distance decreases.

The recognition rate obtained and the quality of the model trajectories depend on the codebook quality and on the model capacity to span the speaker's acoustic space. The configurations used are always realistic, i.e physically reachable by a human vocal tract. This acoustic space is now described for each vowel.

For [i], [œ], [ø]: the third formant values are too low, they never reach the speaker's acoustic space, for [a], [o] and [u]: F3 is too high and never meets the speaker's space; furthermore for [a], F1 is too low and for [i] F2 too high. The model acoustic space matches that of the other vowels for each formant value.

Figure 1 compares the three cardinal vowels [i], [a] and [u] with the DRM model configurations used for speaker TS with the area function proposed by [3].

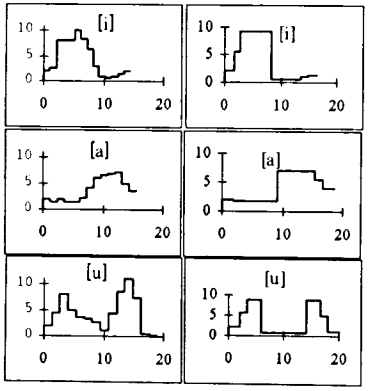

figure 1: Comparison of the three cardinal vowels [i], [a] and [u] with the DRM model configurations (rigth) used for speaker TS with the area function proposed by Majid [3](left). TheX-axis represents the distance from the glottis (cm), the Y-axis represents the regions' areas (cm^2).

Some V1-V2 transitions raise problems. For example [i]-[u] is well recognized when table TT-2-3 is used but the model formant trajectory badly matches the speaker's, the model cannot represent how the F2 and F3 formants cross, which is a characteristic of this transition. Transition [a]-[i] also has a good recognition rate and a poor model representation. For [i]-[y], not only is the third formant of [i] never reached, but also the F2 formant model value is not constant throughout the transition. Model transition with [e] or [ɛ] are acoustically close to the natural curve because these two vowels have good static representations and therefore the total distance decreases.

The transition [ɔ]-[a] always has a good recognition rate when table TT-2-2 and TT-2-3 are used for speakers TS and FB and is always well modelized for each speaker with table TT-2-2 (see figure 2). V1-V2 sequences with [a] also have a good recognition rate.

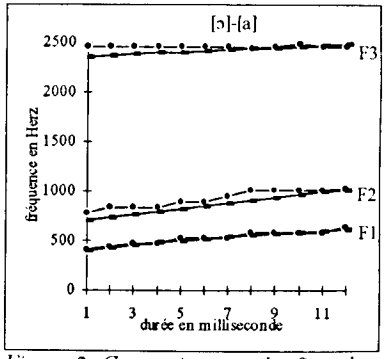

Figure 2: Comparison on the first three formants between the model trajectory ("-") and speaker's ("■ ") for [ɔ]-[a].

3.2 - Dissymetric model

If the symetric axis of the [i] configuration is shifted by 1 cm towards the glottis, the third formant better matches natural vowel space. However, this shift disturbs the [y] acoustic space. This vowel needs the symetric axis to be shifted towards the lips. This leads to change the symetric axis for each vowel and then to change the model characteristics continuously.

The results are not improved if the symetric axis shift is set identically for every vowel. So we keep the initial symetric tube. This shift of the symetric axis of the vocalic tube could be interesting to process a female voice since we know that the characteristics of the female vocal tract are different from those of the male vocal tract.

4 - CONCLUSION

The Distinctive Region Model study leads us to specify the capacities of this production system within the framework of vocal recognition. The constraints we imposed seem to be strong enough to avoid the one-to-many problem. Besides, we need to test different acoustic distances and acoustic parameters to measure the match between the model productions and the speaker's realisations.

Our study shows clearly that speaker adaptation is necessary and significantly improves the results. Moreover, context dependent configurations produce more accurate results.

Some speaker transitions are faithfully represented by the DRM model and have a good recognition rate. However, some particular cases remain to be examined more precisely.

The problem of the acoustic values not reached by the model is not resolved by a dissymetric model. The results are not improved by the use of different interpolation types or the desynchronisation of lip movement.

Generally speaking, our attempt to refine the recognition strategy doesnt improve the model performances. This model is very simple, it cannot respond precisely to any situation but it helps us to obtain suitable results for the vocalic transition recognition. The results obtained, on average 80% of the V1-V2 transition correctly identified, encourage us to carry on speech recognition with the articulatory models. Nevertheless it is advisable to use much more complex models able to take into account the whole articulatory phenomenon (recognition of the articulation mode and place for consonants).

ACKNOWLEDGEMENTS

We would like to thank Cecile Bianchi for her help in translating the article.

REFERENCES

[1] Candille L., H. Méloni, T. Spriet, R. Carré (1994), "Inversion de modèle articulatoire pour la reconnaissance de la parole: application à l'identification de diphones vocaliques",20e JEP, Trégastel, p: 485-490

[2] Carré R. & Mrayati M. (1992), "Distinctive Regions in acoustic tubes. Speech production modeling."Journal d'acoustique 5, 141-159.

[3] Majid R. (1986), "Modélistaion articulatoire du conduit vocal, exploration et exploitation. Fonction de macrosensibilité paramétriques et voyelles du français" Thèse Doc. Ing., INP Grenoble.