3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

VOWEL NORMALIZATION BY ARTICULATORY NORMALIZATION : FIRST ATTEMPS FOR VOWEL TRANSITIONS.

Yohan Payan & Pascal Perrier

Institut de la Communication Parlée - URA CNRS 368 - INPG & Université Stendhal
46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
E-Mail : Perrier@icp.grenet.fr

ABSTRACT

This paper presents a procedure for normalization based on the exploitation of the concept of formant to cavity affiliation as described in the vocalic theory of speech production. For closed vowels, this concept allows an estimation of the ratio of the respective back and front cavity lengths of two speakers, from the ratio of the associated formants. In order to propose a normalization of the vowel space of different speakers we used a "reference speaker", whose articulatory and acoustic properties are well defined : the articulatory model as proposed by Maeda. Vowel normalization consists then in the projection of the vocalic space of one given speaker into the vocalic space of our model. Geometrical relations between the studied speaker and the "reference speaker" are extracted from [i] and [u] for front and velo-palatal vowels, and from [a] and [u] for back and velo-pharyngeal vowels. [y] and [u] give some indications about lip shapes. This method is tested (1) by its ability to predict the standard vowels of a speaker starting from those of the "reference speaker", and (2) by its ability to propose realistic formant patterns for Maeda's model, starting from the corresponding formant transitions of a real speaker.

Keywords : speaker normalization; speaker adaptation; vowel production.

1. INTRODUCTION

Inter-speaker variability is one of the most annoying properties of the speech signal but also one of the most interesting topics for speech research. It is, of course annoying, because it largely contributes to complexify the problem of automatic speech recognition. But it is all the more interesting because it brings up challenging questions about speech production and speech perception: to what extent can sounds be different, or be produced in different ways, and, nevertheless, correspond to the same phoneme? And how is speech perception able to cope with these variable occurrences of the same phoneme?

The answer to these questions implies first the characterization of the variability and the understanding of its causes. Second, it could be very helpful to be able to project the different occurrences of different speakers in a unique space, in order to be able to compare the specific strategies used by different speakers. This corresponds to the problem of speaker normalization, which could be relevant in speech perception process, and which is often included in speech recognition systems.

2. USUAL TOOLS FOR VOWEL NORMALIZATION

In the past years, many works have proposed and tested different approaches for vowel normalization (see [4]). They were based on different signal processing techniques, and different evaluation procedures, but the majority of them included the same basic principles:

① formant extraction

② projection of the variable formant-patterns into a "reference" space; this projection is made in general through linear mathematical transformations, which are parametrized by statistical properties of the acoustic vowel-space of each speaker [5], [7].

From our point of view, this kind of normalization approach, even if it was relatively successfull, is not satisfactory at all, as it is based on a description, and not on an explanation of variability. Therefore, the efficiency of such a method seems essentially limited, as regards normalization possibility as well as contribution to knowledge, and future evolution. In particular, such methods do not refer to any knowledge about the acoustical understanding of formants, and to their articulatory correlates.

In this perspective, the approach adopted by Wakita [9] is more interesting: he attempted to infer some information from the length of the vocal-tract for each speaker with the help of LPC reflection coefficients; knowing that frequency resonances and resonator lengths are correlated, he proposed to normalize all formants for all vowels by the ratio between the vocal-tract length of the speaker and a reference length. But, in reality, the acoustical differences between two speakers can not only be ascribed to vocal-tract length differencies : the anatomical differences are much more complex; Nordström's proposal [8], to apply different length normalizations depending on which part of the vocal-tract is concerned, thus seems to be more suitable.

3. OUR APPROACH

From our point of view, a normalization technic is efficient if it can account for the causes of variability. Acoustical differences are essentially due to two factors: (1) differences in the anatomies of speaker vocal-tracts; (2) differences in speech gestures. Moreover, speech gestures are supposed to produce in a given vocal-tract, with its given anatomy, a required geometry in order to produce the right perceptual effect. It is now often accepted that the vocal-tract geometry is specified in terms of constriction location and geometry, and lip-horn geometry. We suggest then to consider formant differences as essentially related to differences (1) in the front and the back cavities of the vocal-tract, delimited by the oral constriction, and (2) in the lip-horn geometry.

In order to relate these geometrical differences to the corresponding acoustical differences, Fant's four-tube model [3] was adopted as a modelling framework. This model describes explicitely the vocal-tract geometry in terms of lip geometry and of back and front cavities, controlled by the constriction location. It is well-known, that, in general, formants are not uniquely affiliated to one specific cavity of the vocal-tract. But if back and front cavities are weakly coupled, it is correct to exploit the notion of "formant-cavity affiliation", and to associate a given formant with a specific cavity.

In the case of small coupling between cavities, it is thus possible to express the measured formants in terms of:

- — Helmholtz resonance for the set [back cavity + constriction]
- — Half-wavelength resonance in the back-cavity
- — Quater-wavelength resonance in the front cavity for unrounded vowels
- — Half-wavelength resonance in the front cavity for rounded vowels
- — Helmholtz resonance for the set [front cavity + lips] for rounded vowels

4. NORMALIZATION PROCEDURE

Our first assumption is that the consequences of anatomical and gestural differences on the size of back and front cavities could be variable, depending on the location of the articulation in the vocal-tract. We propose therefore to check these differences for vowels describing the whole articulation range: [i], that presents a frontal and high articulation location, [u] that presents a back and high articulation location, and [a] that presents a back and low articulation location. Our second assumption is that the differences in lip gestures will be the same for all rounded vowels. We propose therefore to check them by the analysis of [y] in comparison with [i], because these vowels present in a first appproximation the same articulation location. We propose thus, for the first phase of the normalization procedure, to base our analysis on 4 vowels [i], [y], [u] and [a]. In order to respect the constraint of a small coupling between the cavities, these vowels are produced in a consonantal context, which induces a closure of the constriction : "C'est zizi ça ?"; "C'est zuzu ça ?"; "C'est gougou ça ?"; "C'est rara ça ?".

Starting from Fant's nomograms [3], and taking into account the formant values, we proposed assumptions concerning the formant-cavity affiliation for the speaker.

The second part of our normalization procedure uses Maeda's articulatory model [6] as a "reference speaker" in relation to which the speakers will be normalized. For this purpose, it is necessary to determine, in the model, the articulatory configurations, that correspond to the conditions in which our speaker is analyzed: this operation essentially consists in obtaining closed vowels [i], [y], [u], and [a], with the model, starting from standard vowels already elaborated in this model [2]. Four area-functions are thus generated, from which formant values are calculated. Sensitivity functions around these configurations give the required information about formant-cavity affiliations.

In the third phase, "normalization coefficients" are calculated, in order to give an account of the geometrical differences between the model and the speaker. For vowels [i], [y], [u] and [a], these coefficients are directly obtained from formant values: differences in cavity lengths are accounted for by the ratio of the formants affiliated to these cavities; differences in lip geometry are accounted for by the square root of the ratio of the corresponding Helmholtz resonances. For vowels [e,ɛ,y,ø,œ] which are articulated in the palatal part of the vocal-tract, we assumed that the "normalization coefficients" vary between [i] value and [u] value as a logarithmic function of the distance between the teeth and the constriction location. For [o, ɔ], which are articulated in the pharyngeal part, an identical assumption is made for the variation of the coefficient between [u] and [a].

The last phase of our normalization procedure consists finally in the calculation of the normalized formants, by using the calculated normalization coefficients. After that, we can consider that the normalization procedure is done, and that it would provide results in the acoustic domain for the speaker within the "reference acoustic domain".

5. PREDICTION OF A SPEAKER'S VOWEL SPACE

We now propose a first evaluation of the principles underlying our normalization procedure : it consists in predicting the vowel space of a French speaker, starting from the vowel space of Maeda's model. In a first phase, the normalization coefficients are calculated as described above. Then, the back and front cavity lengths of each vowel of our reference model are modified following the normalization coefficients. Starting from the reference-area-functions generated by Maeda's model, area-functions corresponding to the geometrical characteristics of the speaker are thus obtained. Formants are then calculated by a classical frequency domain model of the vocal-tract [1].

All vowels are produced by the speaker in a favourable consonantal context; Figure 1 and Figure 2 show the data obtained respectively in F1-F2 and F2-F3 planes. Note that the speaker makes no differences between [a] and [ɑ], as well as between [ø] and [œ]. For each vowel, the frequency ranges of the formant frequencies, as measured for several repetitions in the same conditions, is described by the solid contour. The formant patterns obtained with standard configurations of Maeda's model for the same vowels are also plotted (•); in many cases, substantial differences can be observed: for different vowels formants could be outside the frequency range measured on the speaker.

On the contrary, after processing, the predicted formant patterns are much closer to the real patterns of the speaker (Figures 1 and 2): in only one case (vowel [e]), the formants are still outside the frequency range measured on the speaker.

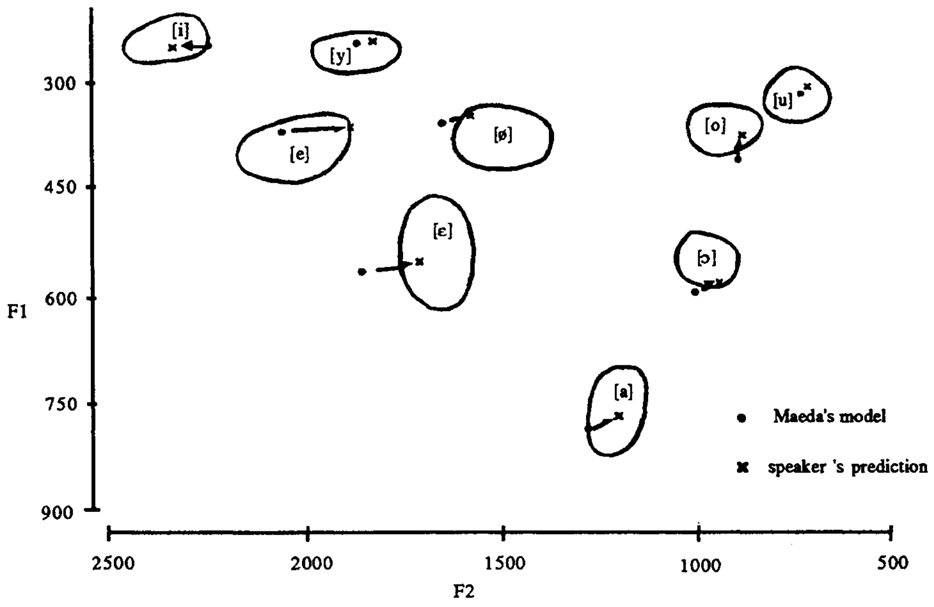

Figure 1: F1-F2 patterns for 9 French vowels as produced by the speaker (frequency range specified by the solid contours), and by Maeda's model (•) and as predicted for the speaker from Maeda's model (+)

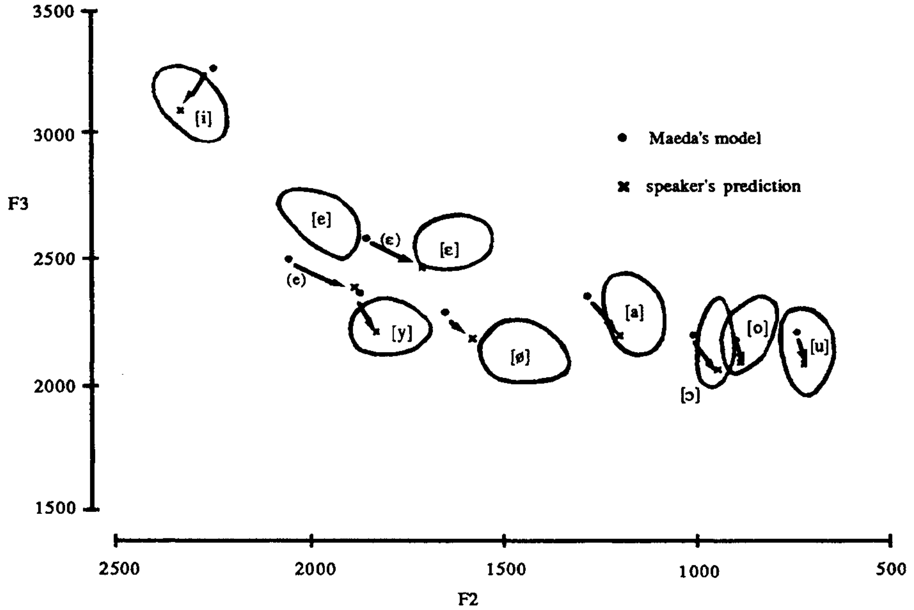

Figure 2: F2-F3 patterns for 9 French vowels as produced by the speaker (frequency range specified by the solid contours), and by Maeda's model (•) and as predicted for the speaker from Maeda's model (+)

6. NORMALIZATION OF VOWEL TRANSITIONS

The precedent results must, of course, be carefully considered, because they only concern one speaker; the question of the choice of the reference configurations on Maeda's model may also arise. In spite of these obvious limitations, it seems correct to assume that useful information about speaker variability can be extracted from the acoustic signal, as long as it is analyzed in terms of vocal-tract resonances, and that this information is relevant for an efficient speaker normalization. Two examples of the normalization procedure, which could thus be used, are now presented.

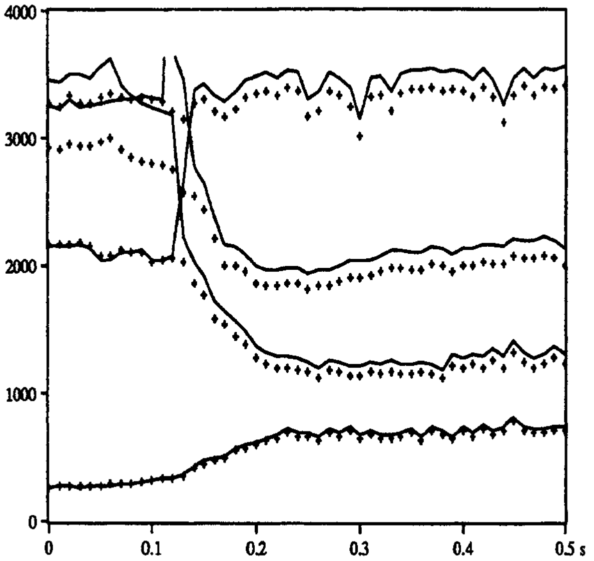

Figure 3: Original formant values (+) and normalized resonances (solid lines) for [iu]

The same speaker pronounced the extreme vowel transitions [iu] and [ia], in the carrier sentence "C'est VV ça ?". Formant frequencies in the transitions are analyzed; they are plotted (+) on figures 3 and 4. The first difficulty consists in the detection of the major affiliation of each formant: the affiliations are actually well-known for [i], [u] and [a], but they are not obvious along transitions. To deal with this problem, basic knowledge of the acoustic theory of vowel production (Fant's nomograms) are used and continuity principles are applied. The estimation of the normalization coefficients along the transition can no more be based on the same approach as in the prediction of the vowel space of the speaker: the constriction location in the vocal-tract is actually not known anymore. However, some information on the variation of this location can be inferred from the resonances: the half-wavelength resonance of the front cavity decreases if the constriction is moved backward; in the same conditions, the Helmholtz resonance of the set [back-cavity + constriction] increases. Figure 3 shows (solid line) the results obtained in the normalization of the transition [iu], where the half-wavelength resonance of the front cavity is used; Figure 4 shows (solid line) the results obtained for [ia], where the Helmholtz resonance is used. In both cases, realistic formant transitions can be observed after normalization; note in particular, how differences between the speaker and the model in the F3 affiliation of [i] can be accounted for: for the speaker, F3 is affiliated to the front cavity, whereas it is affiliated to the back cavity in the model, and this is naturally integrated in the transition.

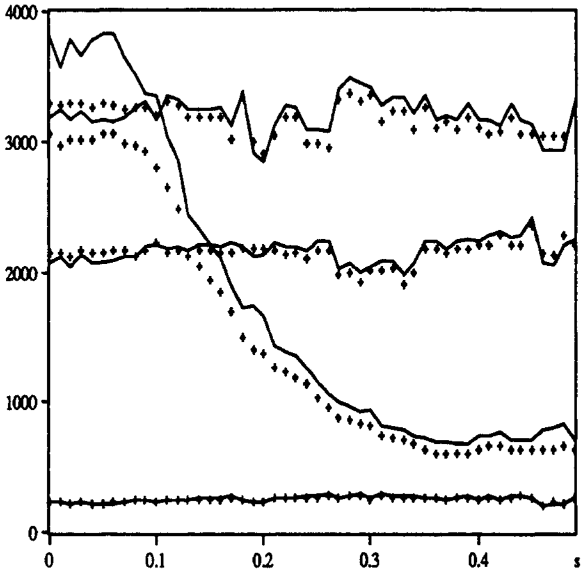

Figure 4: Original formant values (+) and normalized resonances (solid lines) for [ia]

7. CONCLUSION

These preliminary results essentially attempt to illustrate the procedure that we want to develop for a speaker normalization based on the basic principles of the acoustic production of vowels. Prediction of the vowel space of a speaker seems possible, starting from an articulatory model of the vocal-tract, and the normalization of transitions also seems possible following this approach.

Of course, this procedure must be tested on different speakers, and evaluated in comparison with other classical procedures. However, these preliminary observations call for further exploitation of the notion of vocal-tract resonances in a speaker normalization perspective.

REFERENCES

[1]**Badin P. & Fant G. (1984).** Notes on vocal-tract computation. *STL/QPSR 2/3*, 55-108.

[2] **Boë L.J., Perrier P. & Morris A. (1992).** Une prédiction de l'"audibilité" des gestes de la parole à partir d'une modélisation articulatoire. *Proceedings of the 19èmes Journées d'Etude sur la Parole* (pp. 151- 157) Paris: Société Française d'Acoustique.

[3] **Fant G. (1960).** *Acoustic Theory of Speech Production* . The Hague: Mouton.

[4] **Ferrari Disner S. (1980).** Evaluation of vowel normalisation procedures. *J. Acoust. Soc. Am., 67 (1)* , 253-261.

[5] **Lobanov, B.M.(1971).**" Classification of Russian vowels spoken by different speakers," J. Acoust. Soc. Am 49, 606-608.

[6] **Maeda S. (1988).** Improved articulatory model, *J. Acoust. Soc. Am.,81, S146* .

[7] **Nearey, T.(1977).** *Phonetic feature systems for vowels.* Unpublished Doctoral Dissertation, University of Connecticut, Storrs, CT.

[8] **Nordström P-E. (1975).** Attemps to simulate female and infant vocal tracts from male area functions. *STL-QPSR 2-3, 20-33.*

[9] **Wakita H. (1977).** Normalisation of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans. Acoust. Speech Signal Process, ASSP-25*, 183-192.