A model of acoustic interspeaker variability based on the concept of formant–cavity affiliation

Lian Apostol, Pascal Perrier,[a] and Gérard Bailly
Institut de la Communication Parlée, UMR CNRS 5009, INPG, Grenoble, France

(Received 10 December 2002; accepted for publication 20 October 2003)

A method is proposed to model the interspeaker variability of formant patterns for oral vowels. It is assumed that this variability originates in the differences existing among speakers in the respective lengths of their front and back vocal-tract cavities. In order to characterize, from the spectral description of the acoustic speech signal, these vocal-tract differences between speakers, each formant is interpreted, according to the concept of formant–cavity affiliation, as a resonance of a specific vocal-tract cavity. Its frequency can thus be directly related to the corresponding cavity length, and a transformation model can be proposed from a speaker A to a speaker B on the basis of the frequency ratios of the formants corresponding to the same resonances. In order to minimize the number of sounds to be recorded for each speaker in order to carry out this speaker transformation, the frequency ratios are exactly computed only for the three extreme cardinal vowels [i, a, u] and they are approximated for the remaining vowels through an interpolation function. The method is evaluated through its capacity to transform the $(F1,F2)$ formant patterns of eight oral vowels pronounced by five male speakers into the $(F1,F2)$ patterns of the corresponding vowels generated by an articulatory model of the vocal tract. The resulting formant patterns are compared to those provided by normalization techniques published in the literature. The proposed method is found to be efficient, but a number of limitations are also observed and discussed. These limitations can be associated with the formant–cavity affiliation model itself or with a possible influence of speaker-specific vocal-tract geometry in the cross-sectional direction, which the model might not have taken into account. [DOI: 10.1121/1.1631946]

PACS numbers: 43.70.Gr, 43.70.Bk [AL] Pages: 337–351

I. INTRODUCTION

For a long time, the study of variability in speech was mainly guided by speech recognition objectives (Klatt, 1986; Stevens, 1980; Perkell and Klatt, 1986). The aim was thus basically to quantify and to characterize variability in the temporal and spectral domains, not to reproduce it, but to eliminate it, in order to extract *the invariant*, the physical pattern associated with the linguistic input to be recovered. In other words, the aim was to ''normalize'' the speech signal among speakers. For vowels, the purpose of such a technique was to reduce, in the acoustic domain, the variance of the data measured for the same vowel on a number of speakers, in order to enhance the identification scores of the sounds in an automatic classification.

More recently (see in particular Story and Titze, 2002; Titze *et al.*, 1996, 1997; Wong *et al.*, 1996), special attention has been devoted to the generation of interspeaker variability with the aim of contributing to the development of multivoice and multispeaker speech synthesis systems. The present paper is in the line of these studies. Our aim is to propose a speaker transformation technique in the acoustic domain, based on an account of the correlates of acoustic variability in the domain of speaker-specific vocal-tract geometry, with the final objective of predicting the variability in the whole acoustic domain from a limited amount of speech samples for each speaker.

Our basic hypothesis is that the largest amount of interspeaker variability of the formant patterns arises from differences among speakers in the respective lengths of their back and front vocal-tract cavities. These geometrical differences (see, e.g., Bothorel *et al.*, 1986, as an illustration) are due to intrinsic morphological properties of the vocal tract, such as its length (as illustrated by Subjects 3 and 4 in Bothorel *et al.*, 1986), or its distribution between the palatal and the pharyngeal regions (as illustrated by Subjects 1 and 3 in Bothorel *et al.*, 1986), but these differences can also be due to speaker-specific articulatory strategies involved in the positioning and in the displacement of the tongue in the vocal tract (for an example see the [i] production of Subjects 2 and 3 in Bothorel *et al.*, 1986, pp. 16–17). Consequently, in order to transform for each vowel production the formant pattern of speaker A into the formant pattern of speaker B, our approach consists of elaborating a general model for the changes in front and back cavity lengths between these speakers. This model is based on the computation, from the formant frequencies, of back and front cavity length ratios. To do this, it is proposed to apply a basic principle of the acoustic theory of vowel production, namely the formant–cavity affiliation principle. This principle suggests that for a given vowel each formant can be associated, more or less tightly, with a specific cavity of the vocal tract (Chiba and Kajiyama, 1941; Dunn, 1950; Fant, 1960; Stevens, 1972, 1989, 1998), and is therefore essentially determined by the geometry of this cavity.

In this paper, after a short résumé of the main aspects of

[a]Electronic mail: perrier@icp.inpg.fr

this theory and, particularly, of its interpretation in terms of articulatory-to-acoustic relations in vowel production, a method based on these principles and called the *resonance-based method (RBM)* will be presented. A quantitative evaluation of the *RBM* will then be proposed by assessing its capacity to transform the $(F1,F2)$ patterns of five male speakers into the formant patterns of a reference articulatory model of the vocal tract. In order to assess our hypotheses carefully, the corpus we have used consisted of a reduced number of well-controlled logatoms, and, for the purposes of this paper, the variability in connected speech has not been addressed. On the basis of this evaluation, the intrinsic strengths and limits of the concept of affiliation between formants and cavities will be discussed.

II. FORMANTS AND VOCAL-TRACT RESONANCES

A. The formant–vocal-tract cavity affiliation (Fant, 1960)

Fant (1960) has shown that it is possible to obtain fairly good predictions of the formant patterns characterizing oral vowels by using a simple modeling of the vocal tract, consisting of only four tubes. Such a simplification allowed the formants to be specifically interpreted as resonance frequencies of the different cavities of the vocal tract.[1] In this perspective, a basic and efficient tool consists of the well-known nomograms presented by Fant (Fant, 1960, p. 76), which show the variations of the first five formants, when the vocal-tract constriction is shifted from the glottis to the lips: ''*if an advance of the tongue causes a resonance frequency to rise, it can be concluded that the resonance is mainly influenced by a cavity of decreasing length*'' (Fant, 1960, p. 75). This interdependence of resonance and the vocal-tract cavity is all the more evident if the constriction area is small, reducing the acoustical coupling between vocal-tract cavities.

In a vocal tract where cavities are essentially uncoupled, the following resonance modes can be observed (see Fant, 1960, for details):

(i) A half-wavelength resonance mode: its characteristic frequencies are given by the formula

$$R_{\lambda/2}=n\frac{c}{2L}, \tag{1}$$

where c is the sound velocity in the air, L is the length of the considered cavity, and $n \in \mathbf{N}$.

(ii) A quarter-wavelength resonance mode: its characteristic frequencies are given by the formula

$$R_{\lambda/4}=(2n-1)\frac{c}{4L}, \tag{2}$$

with the same notations as above.

(iii) A Helmholtz resonance mode: its resonance frequency is given by the relation

$$R_H=\frac{c}{2\pi}\sqrt{\frac{A}{l\cdot V}}, \tag{3}$$

in which V is the volume of the cavity, and A and l are, respectively, the area and the length of the resonator's ''neck.''

When vocal-tract cavities are acoustically coupled, the relations between geometry and formants do not strictly apply: the more the coupling between cavities, the less the model is valid. In addition, a clear affiliation of formants and cavities happens to be very difficult when the constriction is located in the so-called *focal* regions of the vocal tract. For a vocal tract having its constriction in one of these *focal* regions, the resonance frequencies of the uncoupled cavities have very similar values (see Badin *et al.*, 1990). The size of these regions in the anterior–posterior direction increases with the acoustical coupling between cavities. This phenomenon adds to the difficulty of finding reliable affiliations in the case of a significant acoustical coupling. Note, however, that studying formant variations when the position of the constriction is moved step by step through a focal region of the vocal tract along the back/front direction can help to reduce the uncertainty about affiliations in such a region (see below, Sec. III D).

Having established these two basic alternatives, we should note that reliable formant–cavity affiliations can be hypothesized for a relatively weak coupling between cavities, and outside of the focal regions (Fant, 1960; Mrayati and Carré, 1976; Badin *et al.*, 1990). Fant (1960) and Badin *et al.* (1990) determined these mappings by localizing each vowel on Fant's nomograms so that its ''typical'' formant pattern is produced with plausible constriction position and lip area.

In such conditions, it becomes possible to infer global morphological differences between speakers by simply comparing their respective formants and interpreting them as specific vocal-tract resonances, while assuming that the coupling between cavities for a given vowel is essentially constant among speakers. This is the basic principle of the ''*resonance-based method*'' (called the RBM) that we have formulated to account for interspeaker variability.

B. The resonance-based method

It is known that formant values are influenced by the 3D geometry of the vocal tract. A Helmholtz resonance depends [see Eq. (3)] on the volume of the resonator's ''body,'' as well as on the length and on the cross-sectional area of the resonator's ''neck.'' The half- and quarter-wavelength resonances, which, in theory, depend only on the length (measured parallel to the direction of the air flow) of the associated cavity, are also influenced by the coupling between cavities that depends on the length and on the cross-sectional area of the constriction. Therefore, both differences in cavity lengths and in cavity cross-sectional areas are *a priori* likely to generate interspeaker differences in the acoustic domain. Nevertheless, in the continuity of classical models of interspeaker variability in vowel production published in the literature (e.g., Nordström and Lindblom, 1975; Wakita, 1977; Fant, 1975), for the purposes of this paper we are going to assume that interspeaker differences in cavity lengths are the major factor of interspeaker variability in the formant domain. For formants associated with half-wavelength or quarter-wavelength resonances, this hypothesis means that the influence of interspeaker differences in constriction size, and then in acoustic coupling between cavities, is signifi-

cantly less important than the influence of differences in cavity lengths. This assumption is consistent with the nomograms proposed by Fant (1960) with the four-tube model for two different sections of the constriction tube (see pp. 76–77) and with the horn-shape model for three different minimum constriction areas of the vocal tract (see p. 84). As concerns formants associated with a Helmholtz resonator, the above hypothesis means that the ratio of the body and neck cross-sectional areas has significantly less impact than the ratio between the lengths of these cavities. To our knowledge, this assumption has never been experimentally demonstrated. The present study will contribute to quantitatively evaluate its validity.

According to Fant's nomograms, the number of possible affiliations for the first three formants of an oral vowel is very limited: they are either a quarter-wavelength or a half-wavelength resonance of the front or the back cavity or a Helmholtz resonator of the set ''back cavity+constriction,'' or of the set ''front cavity+lips'' in the case of a rounded vowel (Fant, 1960). Thus, in the framework of this acoustic theory of vowel production, our assumption that interspeaker differences in the area of the vocal-tract constriction can be neglected implies that the variability of the first three formants observed between two speakers *X* and *Y* is explained by the variabilities of the back and front cavity lengths. Therefore, for each vowel v, two length ratios are taken into consideration, one for the front cavity *f*, and the other one for the back cavity *b*

$$\alpha_c[v]=\frac{L_{c,X}[v]}{L_{c,Y}[v]}, \quad c\in\{f,b\}, \tag{4}$$

where $L_{c,X}[v]$ et $L_{c,Y}[v]$ are the lengths of the *c* cavity for speaker *X* and *Y* respectively.

To infer these ratios from the acoustic signal, the relations between resonance frequencies and cavity lengths are used (see equations 1–3). Thus, for half-wavelength as well as for quarter-wavelength resonances, the ratio is expressed as

$$\alpha_c[v]=\frac{L_{c,X}[v]}{L_{c,Y}[v]}=\frac{R_{c,Y}}{R_{c,X}}, \quad c\in\{f,b\}, \tag{5}$$

where $R_{c,X}[v]$ and $R_{c,Y}[v]$ are the lowest half- or quarter-wavelength resonances of the *c* cavity for speaker *X* and *Y*, respectively. For the Helmholtz mode, the length L_c of the resonator's body is not the unique factor determining the resonance frequency R_H. However, according to our hypotheses that interspeaker differences in vocal-tract cross-sectional area and in cavities coupling can be neglected, it can be assumed that the ''shape factor'' $\sqrt{A/l}$ of the resonator's neck is nearly constant, and that differences in volume *V* of the resonator body are mainly due to differences of its length L_c. Under these conditions, the corresponding length ratio can be calculated according to the formula

$$\alpha_c[v]=\frac{L_{c,X}[v]}{L_{c,Y}[v]}\approx\frac{R^2_{H,c,Y}}{R^2_{H,c,X}}, \quad c\in\{f,b\}. \tag{6}$$

In this study, the length ratios will be calculated as formant frequency ratios on the basis of hypotheses concerning the affiliations between formants and cavities. Testing the RBM will allow us to test the validity of these hypotheses.

III. IMPLEMENTATION OF THE RBM FOR SPEAKER TRANSFORMATION IN VOWEL PRODUCTION

A. Selected basis vowels

The RBM was used to elaborate a speaker transformation procedure in vowel production. The aim of this procedure is to be able to generate every vowel sequence of a ''target'' speaker *X* from the recording of the same sequence pronounced by ''source'' speaker *Y*. To do this, the basic idea is to infer general transformation rules from a limited number of vowel sounds recorded both for speaker *X* and *Y*. The choice of these specific sounds is crucial, since they have to carry enough information about the speech articulation of each speaker, so that a generalization concerning the whole vowel space could be possible through our method. For this reason, we decided to use the extreme cardinal vowels [i, u, a]. Indeed, since they correspond in theory to the most extreme speaker-specific articulatory configurations, they effectively cover to a large extent the articulatory space used in vowel production: from [i] to [u], the tongue has a high position in the vocal tract and the lingual constriction moves along the sagittal palatal contour, from the most anterior position to the most posterior one; from [u] to [a], the tongue moves down in the vocal tract from the highest to the lowest position. We'll call these vowels *basis vowels* of the RBM in the rest of this text.

For these *basis vowels* speaker-specific information about the articulatory strategy was extracted through a quantitative comparison with a reference articulatory model of the vocal tract, whose acoustic and articulatory characteristics (i.e., the place of articulation, the area function, and the formants), are precisely known, and whose formant affiliations can be properly determined among the different possibilities suggested by Fant's nomograms.

B. Characteristics of the reference articulatory model of the vocal tract

We used a statistical articulatory model, *Bergame*, developed at the *Institut de la Communication Parlée* in Grenoble on a French speaker (Beautemps *et al.*, 1996) following the method proposed by Maeda (1990). Beautemps *et al.* (1996) performed a ''*statistical analysis of midsagittal vocal tract profiles derived from cineradiographic pictures, recorded in synchrony with video pictures of front views of the lips and with the speech signal.*'' This model generates an area function of the vocal tract from seven articulatory parameters. The associated formant patterns were obtained with the harmonic vocal tract model developed by Badin and Fant (1984).[2]

1. Reference vowels

Reference articulatory configurations were generated with the reference articulatory model of the vocal tract for eight French vowels [i, e, ɛ, a, y, u, o, ɔ]. To do this, two constraints were respected. The first constraint consisted in maintaining inside the whole vowel system realistic articula-

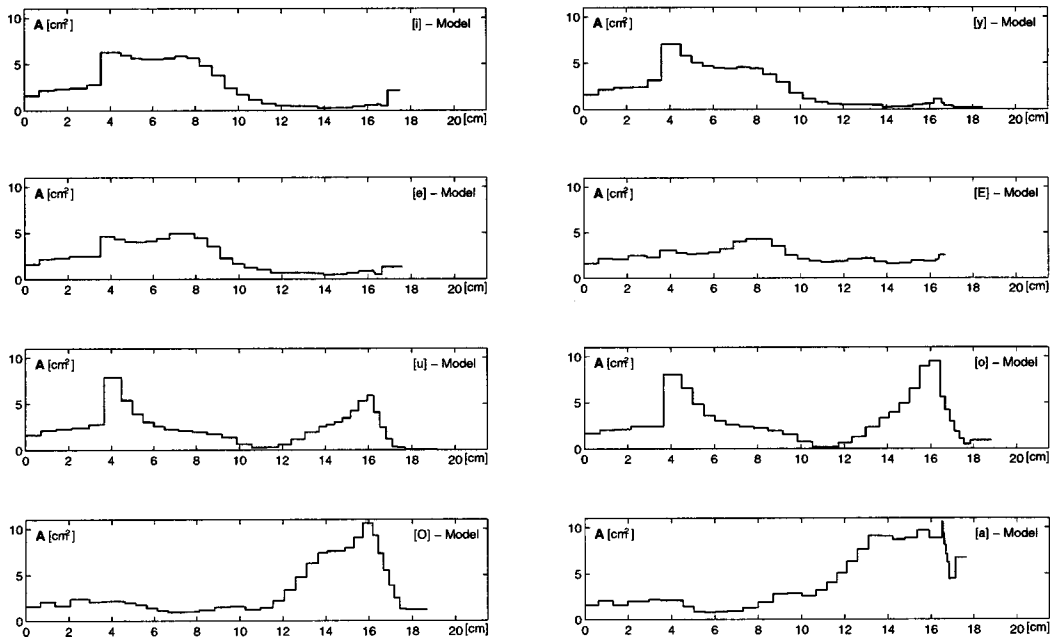

FIG. 1. Area functions obtained for the reference speaker; from the top and from the left to the right: [i, y, e, ɛ, u, o, ɔ, a].

tory positions (Straka, 1965; Abry and Boë, 1986; Bothorel *et al.*, 1986; Majid *et al.*, 1987; Boë *et al.*, 1992): jaw position becomes progressively lower from [i] to [a], passing through [e] and [ɛ]; the lips are more rounded for [y] and [u] than for [o] and [ɔ]; the constriction position in the vocal tract moves back from [i] to [a], passing through [e] and [ɛ], and goes down along the pharynx from [u] to [a], passing through [o] and [ɔ]. The second constraint consisted of ensuring that the formant values of the synthetic reference vowels were close to the ones measured on the corresponding tokens of the French speaker used to develop *Bergame* (Beautemps *et al.*, 1996).

Thus, a set of commands to the articulatory model was obtained for each reference vowel by an acoustic-to-articulatory inversion.[3] A sagittal function, an area function, and the resulting formants were then calculated for each configuration. The eight area functions are given in Fig. 1.

2. Formant–cavity affiliations for the reference vowels

In order to establish the formant–cavity affiliations for the eight reference vowels, our approach was inspired by the vocal-tract sensitivity functions proposed by Fant and Pauli (1974). It consists first of increasing by 10% the cross-sectional area for each elementary section of the area function separately, and, second, in calculating the induced relative formant variations, which we called *sensitivities*. A study carried out on simple tubes, the theoretical resonances of which are known, provided a set of reference patterns for the sensitivities associated with the main resonance modes suggested by Fant's nomograms (1960).

a. Reference sensitivity patterns. Four simple shaped acoustic tubes were chosen for this preliminary study. Each one was obtained by the concatenation of several elementary tubes 0.425 cm in length, so as to reach a total length of 17 cm, which is comparable with the mean length of a male vocal tract (Stevens, 1998). The first tube is closed at one end and has a constant cross-sectional area of 4 cm^2. The first five formants are, in this case, odd multiples of the quarter-wavelength mode of the whole tube ($\lambda/4$, $3\lambda/4$, $5\lambda/4$, etc.). The second tube is similar to the first one, except that it is closed at both ends, its last three sections having a very small cross-sectional area (0.3 cm^2). The first formant of this configuration is therefore a Helmholtz resonance, whereas the next ones are multiples of the half-wavelength mode ($\lambda/2$, λ, $3\lambda/2$, etc.) of the large cavity. The sensitivities associated with the first five formants of these tubes are given in Fig. 2. Two more configurations were obtained by coupling two tubes, which are similar to those described above, with a small tube having a small cross-sectional area. These models are rough representations of a pharyngeal vowel with a large lip opening, and of a rounded velar vowel (see Fig. 3). In the first case, note that the Helmholtz resonance (given by $F1$) concerns only the set "back cavity+constriction," whereas $F2$, $F3$, and $F5$ are the $\lambda/4$, $3\lambda/4$, and $5\lambda/4$ modes of the front cavity, respectively. $F4$ is the half-wavelength resonance of the open–open tube that represents the constriction. In the second case formant sensitivities are those of a couple of Helmholtz resonators.

b. Establishing the formant–cavity affiliations for the reference vowels. The computation of the formant sensitivity curves for the eight area functions of *Bergame* representing the reference vowels permitted for each of them the formulation of the most plausible formant–cavity affiliations. For that, the affiliation was determined first by looking for each formant at the location of the largest sensitivity. Then, the most plausible nature of the resonance (Helmholtz, quarter-wavelength, or half-wavelength) was proposed by comparing the area function and the sensitivity curves with those of the reference patterns. As an example, Fig. 4 shows the formant sensitivities for the vowels [a] (left panel) and [u] (right panel), which are to be compared with the patterns given in

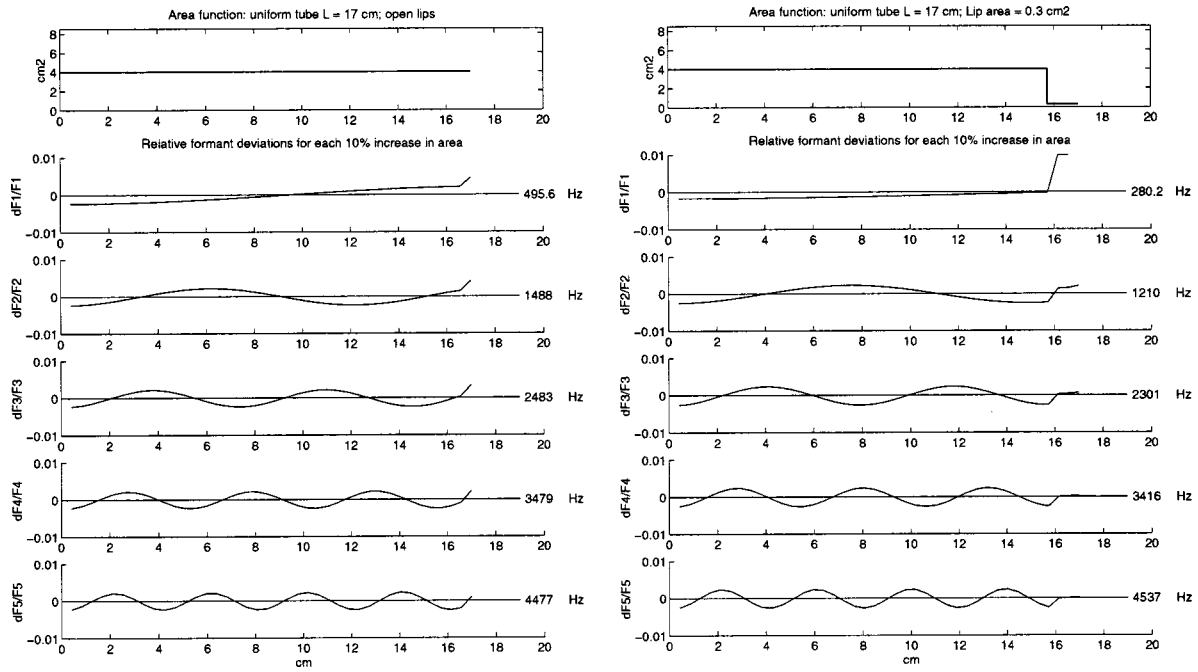

FIG. 2. Formant sensitivities to local area perturbations. Left panel: a closed–open uniform tube. Right panel: a closed–closed uniform tube.

Fig. 3. The affiliations thus established together with the corresponding first four formant values are summarized in Table I.

For all reference vowels but one, the interpretation of the computed sensitivities on the basis of the reference sensitivity patterns led unambiguously to formant–cavity affiliations compatible with a schematization of the area functions with a tube model (see Fig. 3). Vowel [ɛ] is, however, a particular case. It corresponds to an open vocal-tract configuration, without any true constriction separating clearly the vocal tract into two distinct cavities. As a result, the sensitivity curves of this reference vowel (see Fig. 5) are similar to those of the reference pattern depicted in the left panel of Fig. 2: formants are essentially affiliated with the whole vocal tract and they are odd multiples of the quarter-wavelength resonance of the whole vocal tract.

C. Interpolation between basis vowels

As explained above, in our method, the articulatory configurations of the ''target'' speaker *X* and of the ''source'' speaker Y are inferred from the formant patterns of the three

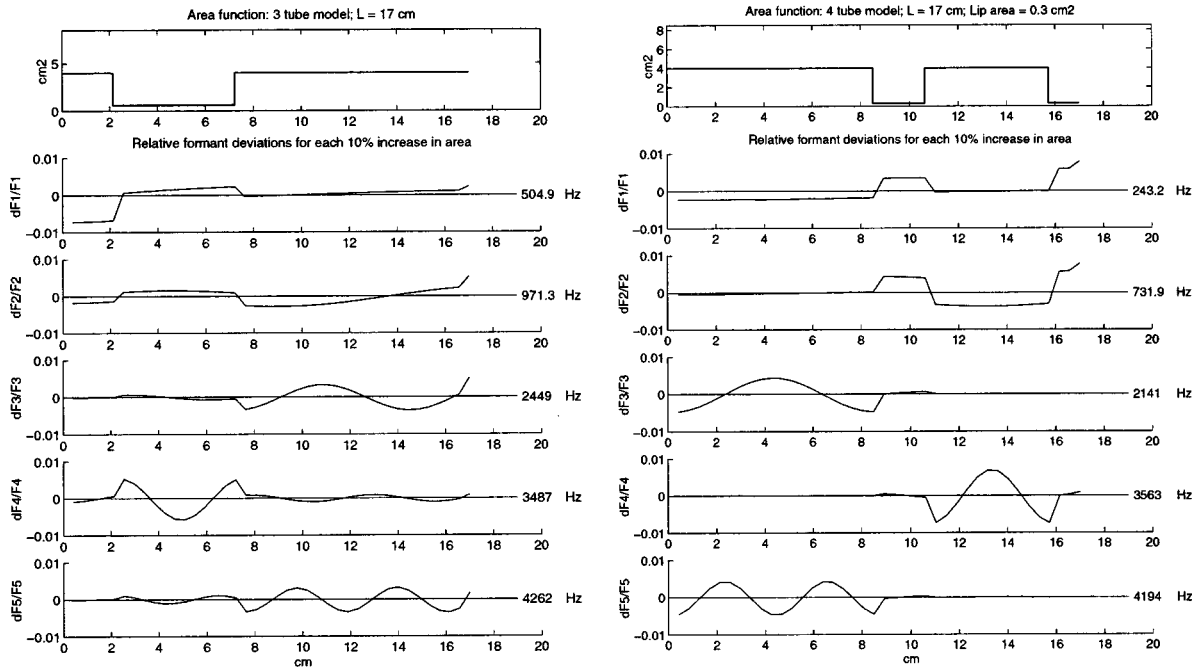

FIG. 3. Formant sensitivities to local area perturbations. Left panel: a tube model simulating a pharyngeal articulation with a large lip opening. Right panel: a tube model simulating a velar articulation with rounded lips.

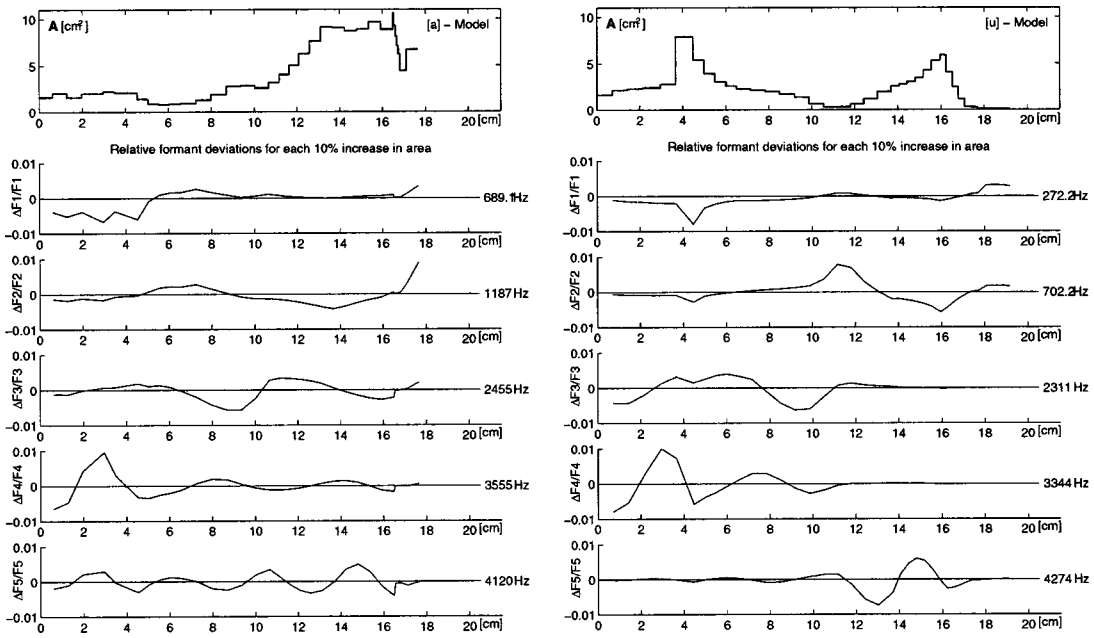

FIG. 4. Formant sensitivities for the reference vowels [a] (left panel) and [u] (right panel) generated with the reference articulatory model of the vocal tract (for comparison see, respectively, the left and right panels of Fig. 3).

basis vowels [a], [i], [u], which give a good account of the speaker's maximal vowel space. These basis vowels are systematically recorded both for the source and for the target speaker. Hence, for the front and for the back cavity, the ratio of the length of each cavity between speakers (α_b for the back cavity and α_f for the front cavity) can be found by the direct computation of the appropriate formant ratios, once the association between formants and cavity resonances is determined (cf. Sec. II B). For the other vowels, which are only recorded for the source speaker, the direct calculation is not possible. Hence, interpolation functions were used to approximate the interspeaker length ratios for these other vowels. For the vowels articulated in the palatal region, like [y], [e], and [ɛ], the coefficients were obtained using a logarithmic interpolation between the ratios of the *basis vowels* [i] and [u]. In a similar way, another logarithmic interpolation between [u] and [a] gave the ratios of the vowels [o] and [ɔ]. The interpolation functions depend on the constriction position measured on the reference vowels (cf. Sec. III A 1).

Thus, for [y], [e], and [ɛ], the ratios were calculated according to the formula

TABLE I. Formant-resonance associations for the vowel prototypes of the *Bergame* model. Legend: *H. back*=Helmohotz resonance of the set ''back cavity+constriction;'' *H. front*=Helmholtz resonance of the set ''front cavity+lips;'' for the other resonances: *back*=back cavity; *front*=front cavity.

Vowel	*F*1	*F*2	*F*3	*F*4
[i]	290	2069	2935	3669
	H. back	λ/2 back	λ/4 front	λ back
[y]	274	1766	2256	3264
	H. back	H. front	λ/2 back	λ back
[e]	349	1932	2641	3638
	H. back	λ/2 back	λ/4 front	λ back
[ɛ]	512	1702	2542	3796
	λ/4	3λ/4	5λ/4	7λ/4
	whole tract	whole tract	whole tract	whole tract
[u]	273	703	2311	3344
	H. back	H. front	λ/2 back	λ back
[o]	340	831	2316	3278
	H. back	H. front	λ/2 back	λ back
[ɔ]	522	911	2364	3370
	H. back	H. front	λ/2 front	λ/2 back
[a]	689	1187	2455	3555
	H. back	λ/4 front	3λ/4 front	λ/2 back

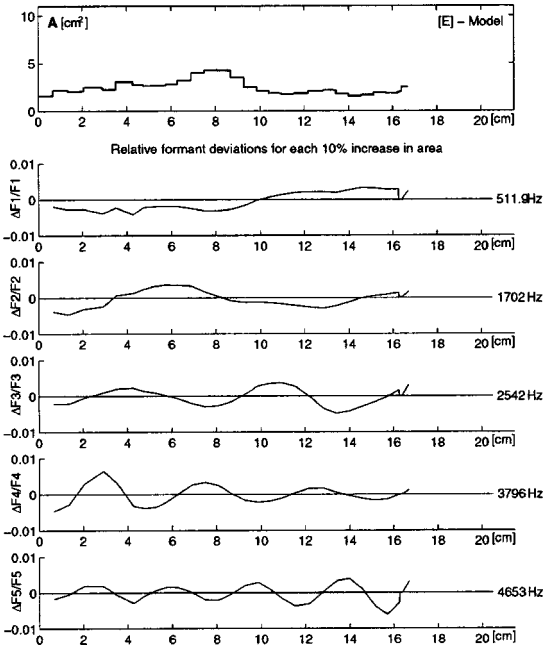

FIG. 5. Formant sensitivities for the reference vowel [ɛ] generated with the reference articulatory model of the vocal tract (for comparison, see Fig. 2, left panel).

$$\alpha_c[v]=\alpha_c[i]+(\alpha_c[u]-\alpha_c[i])\cdot\frac{\log\left(1+CO_c\dfrac{DC[v]-DC[i]}{DC[i]}\right)}{\log\left(1+CO_c\dfrac{DC[u]-DC[i]}{DC[i]}\right)}, \tag{7}$$

and for vowels [o] and [ɔ], α values were obtained from the equation

$$\alpha_c[v]=\alpha_c[a]+(\alpha_c[u]-\alpha_c[a])\cdot\frac{\log\left(1+CP_c\dfrac{DC[a]-DC[v]}{DC[a]}\right)}{\log\left(1+CP_c\dfrac{DC[a]-DC[u]}{DC[a]}\right)}, \tag{8}$$

where $c\in\{f,b\}$, and $DC[v]$ is the distance from the incisors to the center of the constriction for the reference vowel v; CO_c and CP_c are two coefficients that were experimentally determined for each cavity, in order to obtain a good level of efficiency in this procedure.

The calculation of the DC values necessitated the computation of the position of the constriction center for each reference vowel. To do this, the limits of the vocal-tract constriction were determined as the extremities of the vocal-tract's region within which an increase of the cross-sectional area of each elementary tube generates a variation of $F1$ compatible with a Helmholtz resonance: since in this case the constriction is the resonator's neck, $F1$ should increase when the cross-sectional area of the constriction increases (see the positive $\Delta F1/F1$ in the regions of the constriction in Figs. 3 and 4). The position of the constriction center was then determined as the abscissa separating the constriction zone in two parts of equal acoustic impedance. To calculate this abscissa, each elementary acoustic tube was replaced by its low-frequency electrical equivalent (cf. Fant, 1960, p. 28, Fig. 1.2-1). The DC values thus obtained are given in Table II.

D. The approach proposed to determine formant–cavity affiliations for real speakers

Obviously, since the RBM applies to speakers whose vocal-tract geometry is not known, the above approach based on sensitivities of the formants to local geometrical changes could not be used in order to find out the formant–cavity affiliations for real speakers. Therefore, the formant–resonance associations established for the reference vowels were used as initial patterns. However, these relations may not be valid for every speaker, in particular for the three basis vowels, and for vowel [y], because these vowels are located in focal regions of the vocal tract where small changes in vocal-tract geometry can cause changes in the affiliation pattern: this generates an uncertainty about the $F1-F2$ affiliation for [a] and [u]; and about $F2-F3$ or $F3-F4$ affiliation for [i] and [y], (cf. Badin *et al.*, 1990). This is why, in addition to the three *basis* vowels, vowel [y] was also recorded both for the source and for the target speaker, and the affiliation patterns of these four focal vowels were carefully analyzed for each speaker.

For vowel [u], the work of Savariaux *et al.* (1995) on 11 French native speakers showed that $F2$ represents in all cases the Helmholtz resonance of the set "front cavity + lips," whereas $F1$ counts for the Helmholtz resonance of the set "back cavity + constriction." Consequently, we adopted this affiliation pattern for this vowel.

As concerns the other "focal" vowels [i], [a], and [u], there is no such experimental evidence in the literature supporting one affiliation pattern more than another. This is why we recorded for each speaker a number of $V1-V2$ sequences, where $V1$ and $V2$ are either [i], [a], or [u], and analyzed the formant trajectories in these sequences. Indeed, while abrupt changes in formant frequencies can be observed during such sequences when the articulation location goes through a focal point, resonance frequencies vary smoothly and monotonically (Bailly, 1993). Consequently, knowing the articulatory changes involved in each $V1-V2$ sequence, and looking for the corresponding monotonous resonance variations, it was possible to make reliable assumptions about the formant–cavity affiliations.

For example, in an [iy] sequence, while large changes are observed in the lip shape, from spread (for [i]) to rounded (for [y]) lips, the back cavity undergoes basically no modification. Indeed, both the place of articulation and the size of the constriction are similar for the two vowels (cf. Bothorel *et al.*, 1986). Consequently, it is reasonable to consider that the resonance that varies the most during the sequence is affiliated with the front cavity. For [y] the lowest resonance frequency of the front cavity corresponds to $F2$. Therefore, following the variation of the resonance backwards across the sequence allows us to find for [i] which is the lowest formant affiliated to the same cavity. An illustration of the analysis is given in Fig. 6, which presents the formant track-

TABLE II. Incisors-to-center of the constriction distances (in cm) for the vowels of the reference articulatory model of the vocal tract.

Vowel	[i]	[y]	[e]	[ɛ]	[u]	[o]	[ɔ]	[a]
DC [cm]	2.4854	2.7776	2.6030	2.9374	6.0975	6.2873	9.3164	9.9971

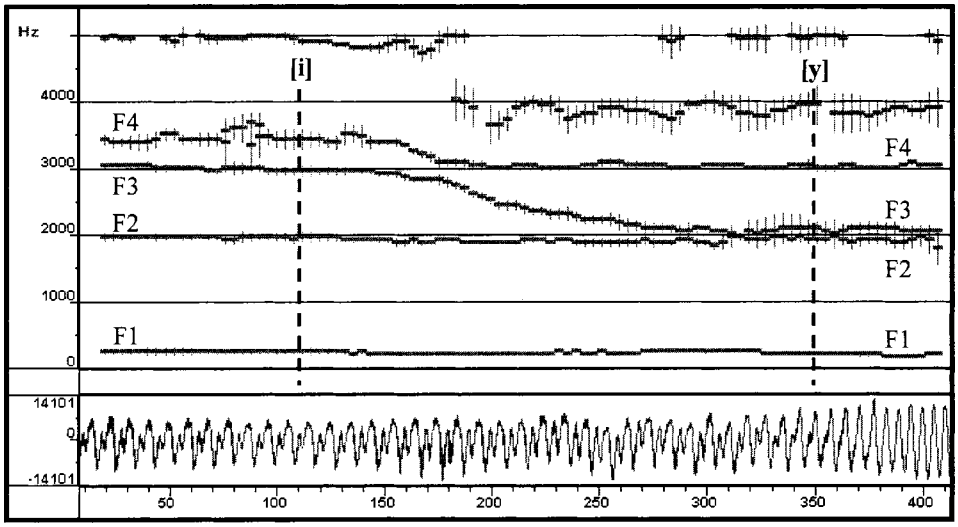

FIG. 6. Formant tracking in the [iy] transition of speaker CS. (Computed with Winsnoori, Babel Technologies, Belgium.)

ing in an [iy] sequence: the first resonance frequency of the front cavity, represented for [y] by $F2$, originates from the fourth formant $F4$ of [i]; consequently, formants $F2$ and $F3$ of [i] are, respectively, the half-wavelength and the wavelength resonance modes of the back cavity, which are in turn represented by $F3$ and $F4$ for [y]. Using a similar approach, it is possible to identify the formant associated with the half-wavelength resonance of the back cavity of [a] by examining the trajectory of the second formant of [i] during an [ia] sequence.

IV. EVALUATION OF THE METHOD

The RBM aims at accounting for interspeaker variability by integrating morphological and articulatory differences between speakers. This method requires us to know which vowel is produced, in order to use the appropriate cavity length ratios. Hence, it is well adapted to speaker transformation purposes, where the objective is to transform a sentence pronounced by a speaker A so that it sounds like a sentence pronounced by speaker B. Interesting studies have been proposed in the last 30 years that contributed to this objective (e.g., Childers *et al.*, 1989; Nordström, 1975, 1977; Story and Titze, 2002; Titze *et al.*, 1996, 1997; Wong *et al.*, 1996). However, for all of them the problem of a quantitative evaluation of their efficiency was never really solved. Therefore, the evaluation framework that we chose for the RBM consisted of transforming five male speakers into the reference articulatory model of the vocal tract, and in measuring the corresponding reduction of the dispersion in the $(F1,F2)$ plane of the formant distribution measured for the whole set of speakers. It permits a quantitative comparison of the RBM with normalization techniques published in the literature.

A. Corpus

Since the reference articulatory model of the vocal tract was built using data from a male speaker without any additional transformation (cf. Sec. III A), we recorded five male speakers. They were all native speakers of French, agreed to participate voluntarily in our experiment, and none of them had any record of pathology of speech production or of the auditory system.

The corpus was designed according to three major requirements for the vowels to be analyzed: (1) reducing the acoustical coupling between the back and front vocal-tract cavities; (2) reducing the token-to-token intraspeaker variability; (3) favoring for each speaker the production of extreme articulations for cardinal vowels [i, a, u]. To this end, the subjects were asked to pronounce the vowels within CVC sequences, where C is a constant, in order to favor the production of the closest possible configuration of vowel V, and thus to reduce the coupling. In addition, consonant C was a voiced consonant, in order to facilitate formant tracking. Each sequence was pronounced within a short word (given in Table III), which was in turn embedded in the carrier sentence: "C'est *CVC* ça?" ("That's *CVC*?"). Subjects were required to repeat each sentence ten times.

In addition to this corpus, the vowel transitions [i–y], [i–a], and [a–i] were also recorded. These transitions were used for each subject to clarify his formant affiliations for the focal vowels [i], [y], and [a] (cf. Sec. III B). They were pronounced in the same carrier sentence as above and each subject repeated them ten times.

The sound recording was carried out in an interactive environment, in an anechoic chamber. The speakers, seating in an armchair, had to read the items displayed on a PC screen. At the beginning of the session, the subjects read the first two phrases of the CVC corpus without being recorded, so that they could get familiar with the environment. Then, for each phrase the acquisition started automatically, as soon as the speech signal level exceeded an experimentally fixed threshold of -24 dB. The speakers were instructed to always favor pronunciation clarity. After the recordings a perceptual verification was performed in order to ensure that the tokens produced by the speakers corresponded well to the desired phonetic category.

The speech signal was captured by a dynamic micro-

TABLE III. CVC[V] items of the closed-context vowel corpus.

Vowel	[i]	[y]	[u]	[e]	[o]	[ɛ]	[ɔ]	[a]
Word	*zizi*	*juju*	*gougou*	*zézé*	*gaugau*	*zêzê*	*troc*	*rara*

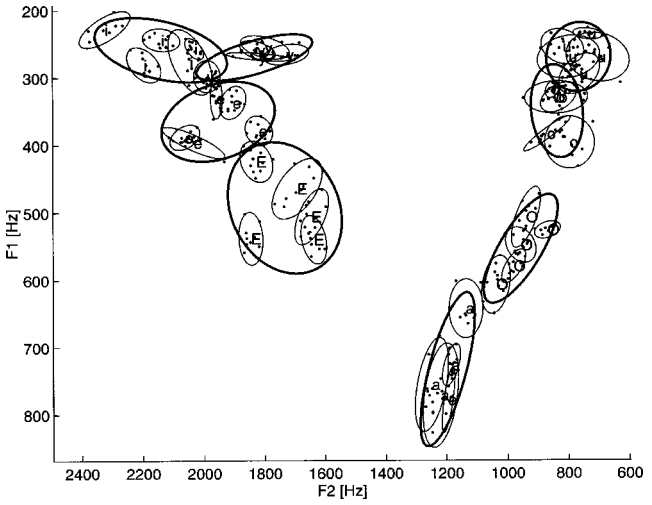

FIG. 7. Vowel distributions in the $(F1, F2)$ plane for the five speakers before transformation using the RBM. The bold line ellipses characterize the dispersions for the whole set of speakers, while individual dispersions are represented with thin line ellipses.

phone [*Beyerdynamic M 10 1N(C)*], amplified by a mixer-amplifier (*Yamaha MV 802*), and sampled at a frequency of 20 000 Hz, using a soundboard (*Oros*) installed on a PC.

Formants were detected with an LPC analysis as follows: 20 coefficients, Hamming window of 20 ms; shift of the window: 5 ms; signal pre-emphasis (coefficient: 0.95). After phoneme labeling, the formants of each vowel were measured in the middle of its steady zone. The formant values used henceforth for the evaluation of the method correspond to the central vowel of the CVC sequences.

Figure 7 shows the vowel distributions for the whole set of speakers in the $(F1,F2)$ plane. The bold line ellipses characterize the dispersions for the whole set of speakers, while individual dispersions are represented with thin line ellipses.

B. Formant affiliations proposed for the real speakers

As explained above, the RBM uses the lowest half-wavelength and quarter-wavelength resonance frequencies or, in some cases, the Helmholtz frequency, in order to infer the ratios of the cavity lengths between two speakers (cf. Sec. II B). The following notations will be used henceforth:

(i) $R1$ is the first resonance mode of the back cavity; usually, it is a Helmholtz resonance of the set ''back cavity+constriction;''
(ii) $R2$ is the lowest resonance of the front cavity, which may be either a quarter-wavelength resonance, if the lips are open, or a Helmholtz resonance, if lips are rounded;
(iii) $R3$ is the second resonance mode of the back cavity, which is in all cases a half-wavelength one.

The formant affiliations proposed for the human speakers are given in Table IV. For [i, a, y] these associations were determined using the vowel transitions (see Sec. III D), whereas for the other vowels the patterns established for the reference vowels (see Sec. III B 2) were taken into consideration. When two affiliations seemed equally possible, we took into account *a posteriori* the affiliation leading to the best results.

TABLE IV. Formant–resonance associations proposed for the vowels of the human speakers. Legend: $R1$ and $R3$ are, respectively, the first and the second resonance modes of the back cavity; $R2$ is the lowest resonance of the front cavity.

Vowel	$R1$	$R2$	$R3$
[a]	$F1$	$F2$	$F4$
[i]	$F1$	$F4$	$F2$
[u]	$F1$	$F2$	$F3$
[o]	$F1$	$F2$	$F3$
[ɔ]	$F1$	$F2$	$F3$
[e]	$F1$	$F3$	$F2$
[y]	$F1$	$F2$	$F3$
[ɛ]	$F1$	$F2$	$F3$

C. Reduction of variability among speakers

The evaluation of the method consisted of two steps. First, the vowel formants of each human speaker were transformed into the formant space of the reference articulatory model of the vocal tract. Then, a comparison was made in the $(F1,F2)$ plane between the scattering of the points before and after transformation. A reduction of this dispersion would attest to the effectiveness of the procedure.

The speaker transformation follows the successive steps of the procedure given in Secs. II B and III C. First, for the back and the front cavity the length ratios between speakers were calculated for the *basis vowels* according to the formulas

$$\alpha_b[v]=R3_M/R3_X, \quad v\in\{\text{a, i, u}\}, \tag{9}$$

$$\alpha_f[v]=R2_M/R2_X, \quad v\in\{\text{a, i}\}, \tag{10}$$

and

$$\alpha_f[u]=\left(\frac{R2_M[u]}{R2_X[u]}\right)^2, \tag{11}$$

where M stands for *reference articulatory Model of the vocal tract* and X for the real speaker to be transformed; $R2$ and $R3$ have the meanings given in Sec. IV B. The $\alpha_f[u]$ expression takes into account the fact that for this vowel, $R2$ is a Helmholtz resonance. The α ratios of the other vowels were determined by interpolation between the respective values of the *basis vowels* (cf. Sec. III C). The ''transformed'' resonances of speaker X were obtained for each vowel as follows:

(i) For a half- or a quarter-wavelength resonance, the transformed value was calculated as the product between the initial value and the α ratio corresponding to the associated cavity.
(ii) For a Helmholtz resonance

$$Ri_N[v]=Ri[v]\sqrt{\alpha_c[v]}, \tag{12}$$

where $(i,c)\in\{(1,b);(2,f)\}$.

The front and back cavity length ratios obtained with this method for the eight vowels and for the five speakers are

TABLE V. Values of the length ratios of the front (F) and back (B) vocal-tract cavities for the eight vowels and the five speakers.

	[i]		[y]		[e]		[ɛ]		[a]		[ɔ]		[o]		[u]	
Speaker	F	B	F	B	F	B	F	B	F	B	F	B	F	B	F	B
CS	0.84	1.02	0.83	1.04	0.84	1.03	0.82	1.04	0.97	0.97	0.92	1.03	0.75	1.07	0.74	1.07
JLS	0.77	0.94	0.78	0.94	0.78	0.94	0.79	0.94	0.95	0.96	0.93	0.94	0.85	0.93	0.85	0.93
MP	0.79	0.97	0.80	1.02	0.79	0.99	0.81	1.03	1	1.	0.98	1.06	0.92	1.1	0.92	1.1
PB	0.75	1.01	0.75	1.02	0.75	1.02	0.76	1.03	1.05	0.98	0.99	1.02	0.79	1.04	0.78	1.04
JMD	0.83	0.89	0.85	0.96	0.84	0.92	0.86	0.97	0.99	1.03	1.	1.06	1.	1.08	1.	1.08

given in Table V. It can be seen that noticeable differences exist for each speaker between the length ratios of the back and front cavities. This supports our strategy of using two length ratios; a unique length factor applied to the whole vocal-tract length would not be able to account for the specific vocal-tract geometry of each speaker. A principal component analysis applied to these data reinforces this statement: two factors are at least required to describe more than 80% of the length ratios variance (the first three factors, respectively, explain 57.6%, 30.6%, and 10.6% of the data variance). In addition, clear differences can be observed between the length ratios obtained for [i], [a], and [u]. This supports our decision to use a specific ratio for each extreme cardinal vowel.

The reduction of variability obtained with the length ratios of Table V was quantitatively measured by calculating for each vowel in the $(F1, F2)$ plane the area of the 2σ dispersion ellipsis, before and after transformation. The results are given in Table VI. Note that the areas are not given in Hz^2, but in percentage. This percentage was calculated as the ratio between the ellipsis area and the area of the rectangle defined by the $F1$ and $F2$ ranges of the global vowel distribution of the five speakers. This was necessary for further comparison with normalization techniques. The dispersion of the vowels obtained for the whole set of speakers after the formants transformation provided by the RBM method is given in Fig. 8. As in Fig. 7, the bold line ellipses characterize the dispersions for the whole set of speakers, while individual dispersions are represented with thin line ellipses.

One can note that except for vowels [o] and [y], the global variability clearly decreases after speaker transformation using the RBM. The clear general reduction of the variability suggests that, in spite of its underlying simplifying hypotheses, the RBM accounts fairly well for the main causes of the interspeaker variability. At the same time, for [y] no noticeable reduction is observed, and for [o] the variability increases clearly. In both cases, the first two formants are Helmholtz resonances. The lack of reduction of the variability could then arise from the fact that focusing on tube lengths variability could be inaccurate, in the cases where, as in Helmholtz resonators, the resonance frequency depends on cavity volumes.

TABLE VI. Areas of the 2σ dispersion ellipses, before and after speaker transformation provided by the RBM. The values are given in percentage with respect to the area of the smallest rectangle including the whole vowel space in the $(F1,F2)$ plane.

	[a]	[i]	[u]	[o]	[ɔ]	[e]	[ɛ]	[y]
Initially	20.90	27.18	15.52	16.85	18.24	30.62	51.90	14.08
RBM	16.77	9.61	8.50	22.06	16.00	12.51	44.95	14.03

Considering the remaining six vowels, it can be observed that the variability decrease after transformation is larger for closed than for open vowels. According to the theoretical hypotheses underlying the RBM, this trend can be explained by the fact that for open vowels interspeaker variability in constriction size and, then, in acoustical coupling between cavities has a non-negligible influence on formant variability.

D. Comparison with speaker normalization techniques

The transformation of the formant patterns of five speakers into a reference formant space can be compared to a classical normalization procedure. Thus, it offers a good framework to evaluate the relevance of the RBM, by comparing the obtained reduction of speaker variability with the ones generated by six normalization techniques published in the literature. Again, it should be emphasized that the RBM cannot be used in a pure normalization framework, since this method requires one to know which vowel is pronounced before it can be transformed into the formant space of another speaker.

Among the considered normalization techniques, four take into account statistical properties of the dispersion of each vowel across speakers. These were proposed by Gerstman (1968), Lobanov (1971), Nearey (1977) (see also

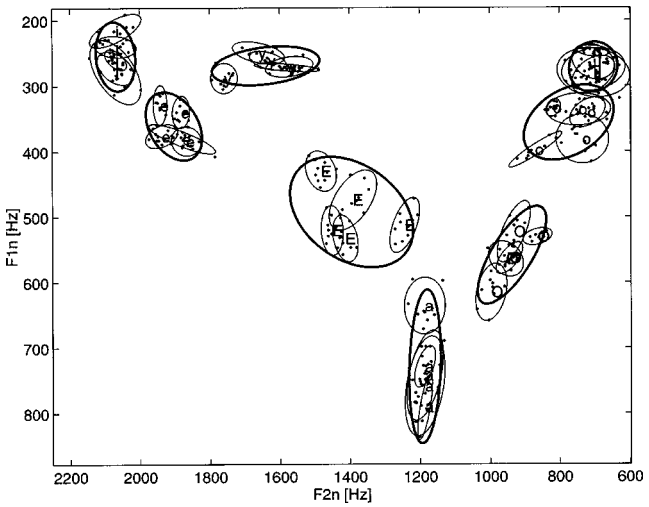

FIG. 8. Vowel distributions in the $(F1, F2)$ plane for the five speakers after transformation using the RBM. See Fig. 7 for legend.

TABLE VII. Areas of the 2σ dispersion ellipses for all the normalization procedures, before and after normalizations. The values are given in percentage with respect to the area of the smallest rectangle including the whole vowel space in the $(F1,F2)$ plane.

	[a]	[i]	[u]	[ɒ]	[ɔ]	[e]	[ɛ]	[y]
Initially	20.90	27.18	15.52	16.85	18.24	30.62	51.90	14.08
Gerstman	18.59	9.38	12.76	23.86	25.82	15.79	74.23	13.15
Lobanov	21.99	11.35	12.88	14.91	21.96	11.95	45.20	14.35
Nearey	12.24	25.33	42.40	29.71	17.06	20.18	34.42	16.64
Di Benedetto	0.00	0.00	0.00	21.14	26.00	20.96	75.31	19.14
Miller (Hz)	21.51	27.68	37.95	43.94	24.89	12.20	30.30	7.20
Miller (Mel)	18.74	22.96	42.14	47.39	23.27	10.81	25.88	7.11
RBM	16.77	9.61	8.50	22.06	16.00	12.51	44.95	14.03

Nearey, 1980), and Di Benedetto and Liénard (1992). The other two techniques were elaborated by Miller (1989). They use nonlinear scale frequencies based on perceptual criteria.

In its original form, Gerstman's (1968) procedure consisted of a homothetic transform that maps the initial variation ranges of $F1$ and $F2$ into a normalized domain, [0, 999] Hz. We implemented a strictly equivalent version of this method (Ferrari Disner, 1980), in which the limits of the normalized $(F1, F2)$ plane are, more realistically, [250, 750] Hz for $F1$, and [850, 2250] Hz for $F2$.

Lobanov's method (1971) normalizes the formant values of each speaker according to the relation

$$F_i^N=(F_i-M_i)/\sigma_i, \quad i\in\{1,2\}, \tag{13}$$

where F_i^N is the normalized value of the F_i formant, and M_i and σ_i are, respectively, the mean value and the standard deviation of F_i calculated for the whole set of vowels.

Nearey (1977) suggested using a logarithmic transformation

$$F_{ijk}^N=\log(F_{ijk})-F_{\text{mean},k}, \tag{14}$$

where i, j, and k represent the formant, the vowel, and the speaker, respectively. $F_{\text{mean},k}$ is the mean of the logarithmic transforms of the first two formants calculated for all the vowels of speaker k.

Di Benedetto and Liénard (1992) proposed to project, through a linear transformation, the $(F1, F2)$ plane of a given speaker, onto a ''standard'' vowel plane defined by the mean values of $F1$ and $F2$ computed for the extreme vowels [a, i, u] produced by a large number of speakers. In concrete terms, for a given speaker, the normalized values of the first two formants of a vowel v, are calculated as follows.

(1) The position of vowel v in the speaker's $(F1, F2)$ plane is expressed as a linear function of the positions of his three extreme vowels [i], [a], and [u] with an *ad hoc* constraint applied to the sum of the weights used in the linear formula. This is expressed by the matrix product

$$\begin{pmatrix} F1[a] & F1[i] & F1[u] \\ F2[a] & F2[i] & F2[u] \\ 1 & 1 & 1 \end{pmatrix}\cdot\begin{pmatrix} \alpha[u] \\ \beta[v] \\ \gamma[v] \end{pmatrix}=\begin{pmatrix} F1[v] \\ F2[v] \\ 1 \end{pmatrix}, \tag{15}$$

summarized by the relation $A\cdot\theta[v]=V$ (16)

(2) Each vowel v in the speaker's $(F1, F2)$ plane is transformed into a normalized vowel v_0 with a linear function. Consequently, the relative position of v_0 in the normalized domain relatively to the extreme vowels [a], [i], and [u] is the same as the position of v in the speaker's domain, and it is possible to write

$$V_0=A_0\cdot\theta[v], \tag{17}$$

where A_0, structured like A, contains the mean values of the first two formants of the vowels [a, i, u] calculated over several speakers.

(3) Since $\theta[v]$ is the same in the two spaces, the normalized $F1$ and $F2$ values for the vowel $[v]$ can be deduced as follows:

$$V_0=A_0\cdot A^{-1}\cdot V. \tag{18}$$

Miller (1989) evaluated the capability of several nonlinear frequency scales to reduce the general interspeaker variability, in particular between groups of men–women–children. The two most efficient transformations were selected for our evaluation, namely $\log_{10}(F2/F1)$ and $\log_{10}(F3/F2)$ expressed first in Hz, and then in Mel.

Table VII allows the comparison of the different techniques described above. The input data were the vowels produced by the real speakers. As in Sec. IV C, the areas of the 2σ dispersion ellipsis, normalized with respect to the size of the considered vowel space, are given. The first line of Table VII gives the initial dispersions of each sound class before normalization. Figures 9–14 illustrate the changes of the

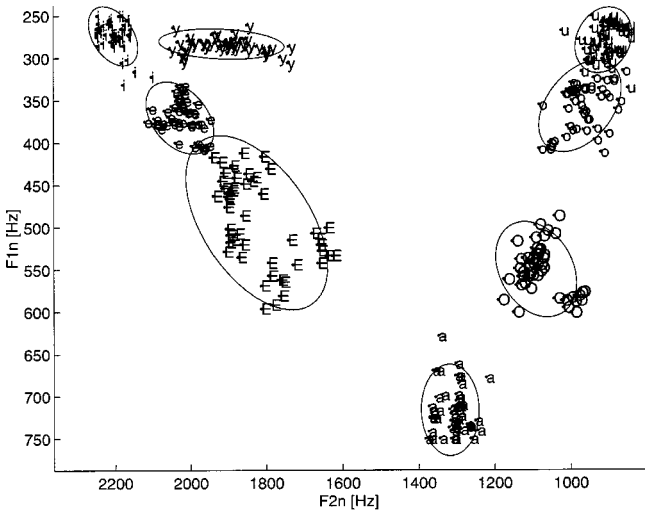

FIG. 9. Vowel distributions in the $(F1. F2)$ plane for the five speakers after normalization with Gerstman's method.

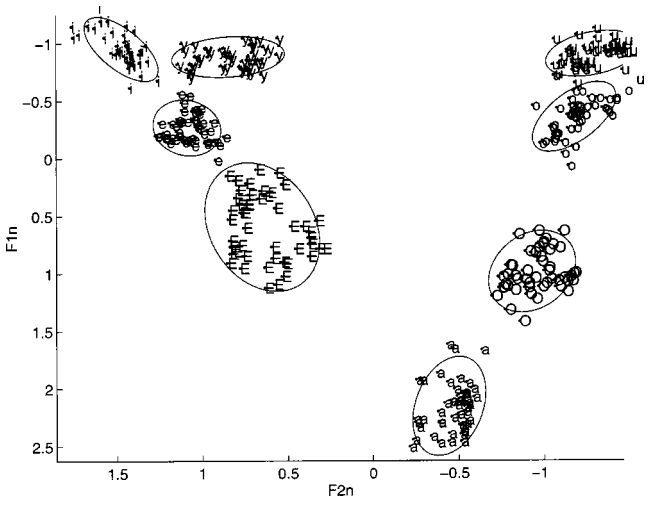

FIG. 10. Vowel distributions in the $(F1, F2)$ plane for the five speakers after normalization with Lobanov's method.

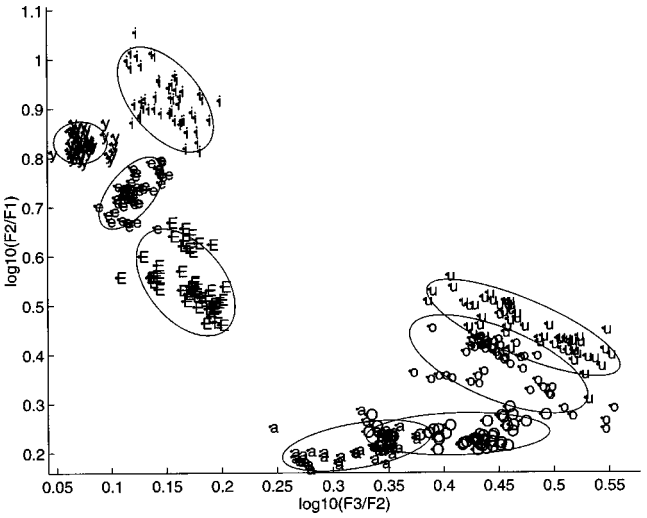

FIG. 13. Vowel distributions in the $(F1, F2)$ plane for the five speakers after normalization with Miller's method in Hz.

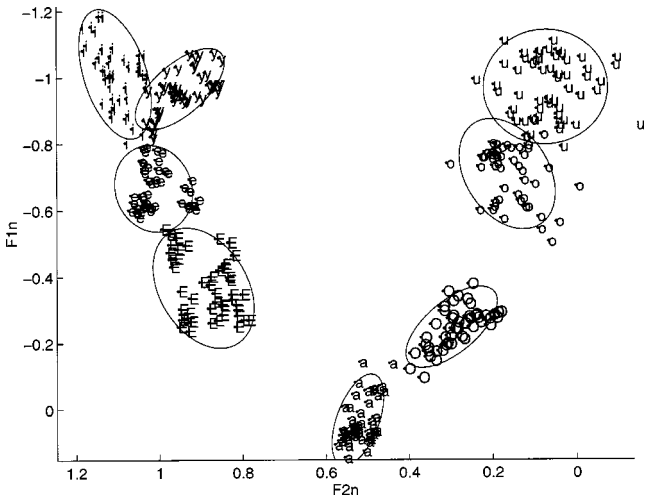

FIG. 11. Vowel distributions in the $(F1, F2)$ plane for the five speakers after normalization with Nearey's method.

acoustic space after application of each of the normalization methods. They can be compared with Fig. 8, which shows the results of the transformation with the RBM. The RBM is, thus, not systematically the most successful procedure, but in almost all cases it is one of the most efficient methods. A more quantitative assessment was made by establishing a ranking of the different procedures computed on the basis of their average ellipsis area. It should be noted that the rankings established for [a], [i], and [u] do not take into account Benedetto and Liénard's procedure, since it intrinsically eliminates the variability for these extreme vowels. The results are given in Table VIII. It can be seen that the RBM is the method that generates on average the largest variance reduction. This result is especially positive, since, contrary to the normalization methods, the transformation provided by the RBM is based on information taken from only three vowels, the *basis* vowels, and considering only the first three formants.

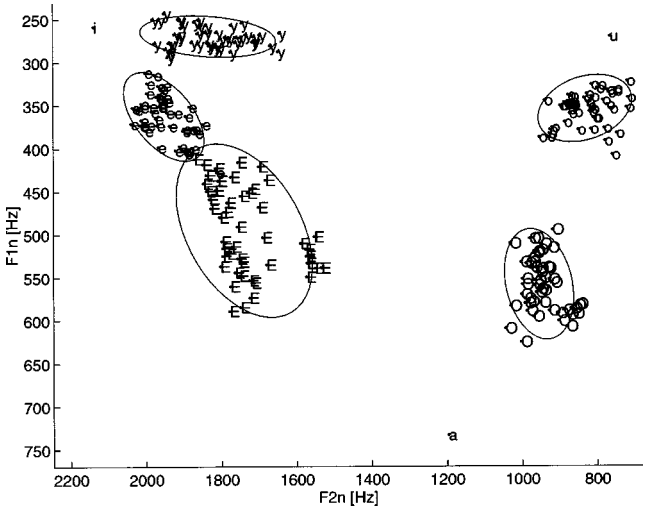

FIG. 12. Vowel distributions in the $(F1, F2)$ plane for the five speakers after normalization with Di Benedetto and Liénard's method.

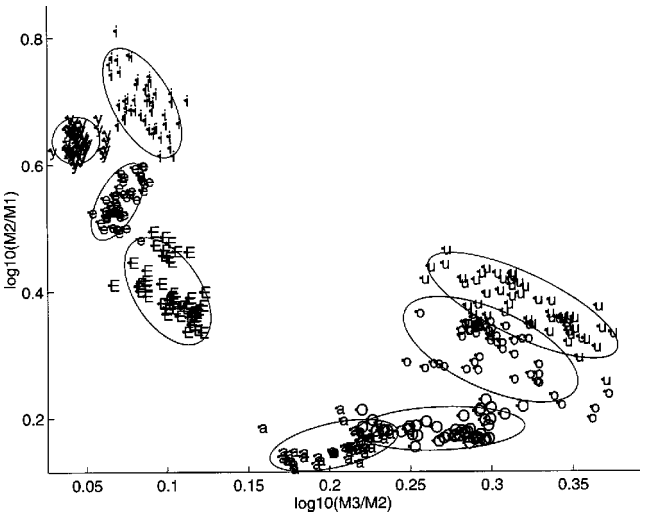

FIG. 14. Vowel distributions in the $(F1, F2)$ plane for the five speakers after normalization with Miller's method in Mel.

TABLE VIII. Average ellipsis areas computed for all vowels (for Di Benedetto and Lienard's method see the text).

RBM	Lobanov	Gerstman	Nearey	Miller (Mel)	Miller (Hz)	Di Benedetto
18.054	19.324	24.198	24.748	24.788	25.709	32.51

V. DISCUSSION

The first conclusion of our study is that interpreting the relations between vocal-tract geometry and formant frequencies by using the formant–cavity affiliation concept is an efficient approach to give a good account of interspeaker variability. These results support our hypothesis that considering separately the length of the back and of the front vocal-tract cavities is an interesting method to model the origins of interspeaker variability. However, beyond this first positive statement, the detailed comparison of Figs. 7 and 8 shows that some aspects of the results obtained with the RBM are not fully satisfactory.

A. Limitations of the speaker transformation based on the RBM

Comparing Figs. 7 and 8, it appears that for vowels [i], [e], [ɛ], [ɔ], and [a] the decrease of the dispersion in the $(F1, F2)$ plane is mainly due to a reduction of the $F2$ variability range, while no noticeable change is observed along the $F1$ axis. For these five vowels, the clear decrease of the $F2$ variability suggests that the affiliations and the corresponding values of the length ratios proposed in the RBM are correct. Now, how can we explain the very weak impact of our speaker transformation procedure on $F1$? Two explanations can be proposed. The RBM assumes that $F1$ is a Helmholtz resonance of the set "back cavity + constriction." Therefore, a first possible explanation would be that, contrary to our assumption, changes in the cross-sectional plane would contribute significantly to the variation of $F1$. In such a case, modeling changes in cavity length would only partially account for the interspeaker variability of $F1$. An alternative explanation could be that, at least for vowel [a], $F1$ would not be a Helmholtz resonance (see below for further analysis).

For vowel [u], Fig. 8 shows a clear decrease of the variability both in $F1$ and $F2$ directions, but this is not the case for the other rounded vowels [y] and [o], especially along the $F1$ axis. Here again, two possible explanations can be proposed. As said above, since for these two vowels $F1$ and $F2$ are considered to be Helmholtz resonances, taking into account changes in the cross-sectional plane in order to integrate volume changes could be a necessity in order to model the interspeaker variability with enough accuracy. An alternative or additional explanation could be that the association between formants and resonances was not correctly established for these two vowels (see below for further analysis).

Thus, in addition to the already mentioned possible non-negligible coupling between cavities, it appears that errors in the association between formants and resonances could contribute to the limitation of the efficiency of the RBM. These errors can be due to an inadequate acoustic model for vowel production or to difficulties in establishing the correct affiliations in the framework of Fant's four-tube model.

B. An alternative acoustic model for some vowels?

Fant's four-tube model and its associated vocal-tract nomograms, on which we based on RBM, do constitute a generally well-accepted reference for the understanding of vowel acoustics. Nevertheless, alternative well-known models have also been proposed in the literature, which would lead to somewhat different hypotheses. In particular, Stevens (1972, 1989, 1998) suggested modeling the vowel area functions only with three tubes and even with two tubes, when the vowel is articulated either in the very back or the very front region of the vocal tract. Thus, for vowel [a], Stevens' model consists of the concatenation of a narrow tube accounting for the constriction, and of a larger tube representing the front cavity (see Stevens, 1998, p. 274). For very front vowels like [i], Stevens (1989, 1998) suggested using a three-tube model (see Stevens, 1998, p. 277). Consequently, both for [a] and for [i], the affiliation patterns suggested by Stevens are different from those that were taken into consideration in the RBM on the basis of the four-tube model. The affiliations suggested by Stevens for the first four formants of [i] and [a] are given in Table IX.

Thus, for [a], the lowest resonance affiliated with the back cavity would be a quarter-wavelength resonance instead of a Helmholtz resonance (see also Mrayati and Carré, 1976; Bailly, 1993). Considering this suggestion for some speakers in the RBM could possibly enhance the decrease of the $F1$ range after transformation for [a].

For vowel [i], Stevens' three-tube modeling is associated with a shortening of the front cavity and a lengthening of the constriction (Stevens, 1998, p. 277). In this case, the third formant would not be a quarter-wavelength resonance of the front cavity, but rather a half-wavelength resonance of the constriction tube. Here again, it would be interesting to observe the consequences of this suggestion on the RBM.

Both kinds of models are generally considered to be plausible.[4]

C. Inaccuracy of the models in focal regions

Another limitation of the RBM arises from the fact that the extreme vowels that serve as *basis vowels* in our method

TABLE IX. Association between formants and vocal tract resonances for [a] and [i] as suggested by Stevens' model (1989, 1998), Legend: *H. back* = Helmholtz resonance of the set "back cavity + constriction;" for the other resonances: *back* = back cavity; *front* = front cavity.

Vowel	$F1$	$F2$	$F3$	$F4$
[a]	λ/4 front	λ/4 back	3λ/4 front	3λ/4 back
[i]	H. back	λ/2 back	λ/2 constriction	λ back

are articulated in focal regions of the vocal tract. Such regions appear both on the nomograms generated with the four-tube model (Fant, 1960; Badin *et al.*, 1990), and on those obtained with the two- and/or three-tube model (Stevens, 1972, 1989, 1998). In the case of a two-tube model, focal points correspond to the merging of the quarter-wavelength resonances of the front and of the back cavities. In the case of three and four tubes, if the constriction tube is long enough, a focal region can correspond to the merging of up to three resonances. Therefore, for a configuration articulated in such a region, there is an uncertainty concerning the affiliations between formants and cavities, even if the coupling between cavities is very small. When the coupling increases, the uncertainty zone widens on both sides of the focal point.[5] For a vowel articulated in such a region, affiliations can be properly determined only if one knows on what side of the focal point the vowel is.

Given the interpolation functions that are used in the RBM, errors in formants–cavities affiliation patterns for *basis* vowels could generate errors in the whole $(F1, F2)$ plane. Therefore, looking for possibilities to increase the accuracy of the determination of these patterns would noticeably increase the reliability of the RBM.

VI. CONCLUSION

Our original hypothesis was that interspeaker variability of the formant patterns for oral vowels originates in differences in the lengths of the back and front vocal-tract cavities, and that it is possible to infer them from an analysis of the formants in terms of vocal-tract cavity resonances. Our results support this assumption, since the *resonance-based method* permits us to give a fair account of the specificity of each speaker's speech production. In particular, the generally good reduction of the variability along the $F2$ axis supports the validity of the concept of affiliation between formants and cavities, as well as of the majority of the affiliations that were chosen in the procedure. It also supports the hypothesis that the influence of interspeaker differences in acoustical coupling between cavities can be neglected as compared to the influence of differences in cavity lengths.

However, limitations of the method were also shown, which could be linked either to an inexact determination of some affiliations, especially in the focal regions, or to the fact that interspeaker geometrical differences in the cross-sectional plane cannot be neglected. Considering an alternative model of vowel production, such as Stevens' (1998) model, could contribute to solve at least a part of the inappropriate hypotheses about the affiliations. Collecting 3D geometrical data on a number of speakers and studying the associated acoustic signal would permit us to evaluate quantitatively the influence on the acoustics of the variability in the cross-sectional plane.

In general, the RBM was shown to be an interesting tool to understand and model interspeaker variability in vowel production, when the subjects are recorded in laboratory conditions and in a consonantal context that favors the most prototypical articulation of the vowels. Working on connected speech, which is usually hypoarticulated, and for which more acoustical coupling between cavities is to be expected, should permit us to assess the practical usefulness of the RBM for speech technology.

ACKNOWLEDGMENTS

This work was supported by doctoral scholarships from the *Robert Schuman Foundation* and from the *Région Rhônes-Alpes (Program TEMPRA)* to the first author.

[1]The idea that a formant may be considered as a characteristic resonance of a vocal-tract cavity had already been suggested (cf. Chiba and Kajiyama, 1941; Dunn, 1950).

[2]This model takes into account all the boundary conditions: the heat conduction and the viscosity losses are considered with unitary shape factor; the radiance at the lips is modeled by a piston in an infinite plane, and the wall vibrations are accounted for by localized impedances along the vocal tract, independently of the area function. There is no subglottal coupling.

[3]This inversion is based on a gradient technique that has systematically as an initial condition the neutral shape of the vocal tract (for more details, cf. Bailly *et al.*, 1995).

[4]Note that Fant himself sometimes used a two-tube modeling, in particular for [i] (e.g., Fant, 1966, 1975).

[5]Indeed, the more the coupling is important, the more the slope of the formants' trajectories moves away from that of the resonances of the decoupled cavities, and comes close to zero (cf., e.g., Fant, 1960, p. 77).

Abry, C., and Boë, L.-J. (**1986**). "Laws for lips," Speech Commun. **5**, 97–104.

Badin, P., and Fant, G. (**1984**). "Notes on vocal tract computations," *Speech Transmission Laboratory—Quarterly Progress and Status Report* (Royal Institute of Technology, Stockholm), **2–3**, pp. 53–108.

Badin, P., Perrier, P., Boë, L.-J., and Abry, C. (**1990**). "Vocalic nomograms: Acoustic and articulatory considerations upon formant convergence," J. Acoust. Soc. Am. **87**(3), 1290–1300.

Bailly, G. (**1993**). "Resonances as possible representation of speech in the auditory-to-articulatory transform," in Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech 93), Berlin, Germany, Vol. 2, pp. 1511–1514.

Bailly, G., Boë, L.-J., Vallée, N., and Badin, P. (**1995**). "Articulatory-acoustic vowel prototypes for speech production," in Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech 95), Madrid, Spain, Vol. 3, pp. 1913–1916.

Beautemps, D., Badin, P., Bailly, G., Galván, A., and Laboissière, R. (**1996**). "Evaluation of an articulatory-acoustic model based on a reference subject," 1st ESCA Tutorial and Research Workshop on Speech Production Modelling—4th Speech Production Seminar, pp. 45–48.

Boë, L.-J., Perrier, P., and Bailly, G. (**1992**). "The geometric vocal tract variables controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion," J. Phonetics **20**, 27–38.

Bothorel, A., Simon, P., Wioland, F., and Zerling, J.-P. (**1986**). *Cinéradiographie des Voyelles et des Consonnes du Français (Cineradiography of French Vowels and Consonants)* (Travaux de l'Institut de Phonétique de Strasbourg: Université des Sciences Humaines, Strasbourg, France).

Chiba, T., and Kajiyama, M. (**1941**) (reprinted in 1958). *The Vowel—Its Nature and Structure* (Tokyo, Japan).

Childers, D. G., Wu, K., Hicks, D. M., and Yegnanarayana, B. (**1989**). "Voice conversion," Speech Commun. **8**(2), 147–158.

Di Benedetto, M. G., and Liénard, J.-S. (**1992**). "Extrinsic normalization of vowel formant values based on cardinal vowels mapping," in Proceedings of the 2nd International Conference on Spoken Language Processing, Banff–Alberta, Canada, pp. 579–582.

Dunn, H. K. (**1950**). "The calculation of vowel resonances, and an electrical vocal tract," J. Acoust. Soc. Am. **22**, 740–753. [Reprinted in: Flanagan, J. L., and Rabiner, L. R., editors (**1973**). *Speech Synthesis* (Dowden, Hutchinson, and Ross, Stroudsburg, PA.]

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands).

Fant, G. (**1966**). "A note on vocal tract size factors and non-uniform F-pattern scalings," *Speech Transmission Laboratory—Quarterly Progress and Status Report* (Royal Institute of Technology, Stockholm), **4**, pp. 22–30.

Fant, G. (**1975**). ''Non-uniform vowel normalization,'' *Speech Transmission Laboratory—Quarterly Progress and Status Report* (Royal Institute of Technology, Stockholm), **2–3**, pp. 1–19.
Fant, G., and Pauli, S. (**1974**). ''Spatial characteristics of vocal tract resonance modes,'' Speech Commun. Seminar **2**, 121–132.
Ferrari Disner, S. (**1980**). ''Evaluation of vowel normalization procedures,'' J. Acoust. Soc. Am. **67**(1), 253–261.
Gerstman, L. J. (**1968**). ''Classification of self-normalized vowels,'' IEEE Trans. Audio Electroacoust. **AU-16**, No. 1, 78–80.
Klatt, D. (**1986**). ''The problem of variability in speech recognition and in models of speech perception,'' in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 300–319.
Lobanov, B. M. (**1971**). ''Classification of Russian vowels spoken by different speakers,'' J. Acoust. Soc. Am. **49**(2), 606–608.
Maeda, S. (**1990**). ''Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model,'' in *Speech Production and Speech Modelling*, edited by W. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 131–149.
Majid, R., Abry, C., Boë, L.-J., and Perrier, P. (**1987**). ''Contribution à la classification articulatori-acoustique des voyelles: Étude des macrosensibilités à l'aide d'un modèle articulatoire'' (''A contribution to vowel classification in the articulatory and acoustic domains: Studying macrosensitivities with an articulatory model''), in *Proceedings of the 11th International Congress of Phonetic Sciences* (Tallin, Estonia), Vol. 2, pp. 348–351.
Miller, J. D. (**1989**). ''Auditory-perceptual interpretation of the vowel,'' J. Acoust. Soc. Am. **85**(5), 2114–2134.
Mrayati, M., and Carré, R. (**1976**). ''Relations entre la forme du conduit vocal et les caractéristiques acoustiques des voyelles françaises—Etude des distributions spatiales'' (''Relations between vocal tract geometry and acoustic characteristics for French vowels—A study of spatial distributions''), Phonetica **33**, 285–306.
Nearey, T. (**1977**). ''Phonetic feature systems for vowels,'' Doctoral dissertation, University of Connecticut, Storrs, CT. Reproduced by Indiana University Linguistics Club, 1978. Unavailable. Cited by Ferrari Disner (1980).
Nearey, T. (**1980**). ''On the physical interpretation of vowel quality: Cinefluorographic and acoustic evidence,'' J. Phonetics **8**, 213–241.
Nordström, P. E. (**1975**). ''Attempts to simulate female and infant vocal tracts from male area functions,'' *Speech Transmission Laboratory—Quarterly Progress and Status Report* (Royal Institute of Technology, Stockholm), **2–3**, pp. 20–33.
Nordström, P. E. (**1977**). ''Female and infant vocal tracts simulated from male area functions,'' J. Phonetics **5**, 81–92.
Nordström, P. E., and Lindblom, B. (**1975**). ''A normalization procedure for vowel formant data,'' in Proceedings of the 8th International Congress of Phonetic Sciences in Leeds, Paper 212.
Perkell, J. S., and Klatt, D. H., (editors) (**1986**). *Invariance and Variability in Speech Process* (Erlbaum, Hillsdale, NJ).
Savariaux, C., Perrier, P., and Orliaguet, J. P. (**1995**). ''Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production,'' J. Acoust. Soc. Am. **98**(5), Pt. 1, 2428–2442.
Stevens, K. N. (**1972**). ''The quantal nature of speech: Evidence from articulatory-acoustic data,'' in *Human Communication: A Unified View*, edited by E. E. David, Jr. and P. B. Denes, McGraw-Hill, New York), pp. 51–66.
Stevens, K. N. (**1980**). ''Acoustic correlates of some phonetic categories,'' J. Acoust. Soc. Am. **68**(3), 836–842.
Stevens, K. N. (**1989**). ''On the quantal nature of speech,'' J. Phonetics **17**, 3–45.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (The MIT Press, Cambridge, MA).
Story, B. H., and Titze, I. R. (**2002**). ''A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function,'' J. Phonetics **30**(3), 485–509.
Straka, G. (**1965**). *Album Phonétique* (*Phonetic Album*) (Presses de l'Université de Laval, Québec, Canada).
Titze, I., Wong, D., Story, B., and Long, R. (**1996**). ''Voice transformation with physiologic scaling principles,'' NCVS Status and Progress Report, **10**, pp. 103–110.
Titze, I., Wong, D., Story, B., and Long, R. (**1997**). ''Considerations in voice transformation with physiologic scaling principles,'' Speech Commun. **22**, 113–123.
Wakita, H. (**1977**). ''Normalization of vowels by vocal-tract length and its application to vowel identification,'' IEEE Trans. Acoust., Speech, Signal Process. **25**(2), 183–192.
Wong, D., Lange, R. C., Long, R. K., Story, B. H., and Titze, I. R. (**1996**). ''Age and gender related speech transformations using linear predictive coding,'' NCVS Status and Progress Report, **10**, pp. 111–126.