



ELSEVIER

Speech Communication 16 (1995) 27–47

SPEECH
COMMUNICATION

Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data ^{*}

Denis Beautemps ^{*}, Pierre Badin, Rafael Laboissière

Institut de la Communication Parlée, URA CNRS N°368, INPG – Université Stendhal, 46 Avenue Félix Viallet, 38031 Grenoble Cedex, France

Received 20 April 1994; revised 8 August 1994

Abstract

In order to achieve better understanding of the articulatory–acoustic relationships, more data are still very much needed. The two-fold aim of the present study was thus (1) to provide a set of coherent midsagittal functions, area functions and formant frequencies, for a small corpus of vowels and fricative consonants produced by one subject, and (2) to derive a midsagittal profile to area function conversion model optimised for this given subject. Simultaneous tomography and sound recording were available for the subject, as well as some complementary data such as lip geometry or casts of the hard palate. The model is based on Heinz and Stevens' $A = \alpha d^\beta$ area function model, modified so that α varies continuously along the vocal tract midline as a function of the midsagittal distance. The coefficients of the model have been determined with the help of an optimisation algorithm based on a gradient descent technique. The gradient of the error between actual and desired formant values was computed through a back-propagation network implementing both sagittal-to-area conversion and acoustic wave propagation. The fact that the model should work for sounds as different as vowels and consonants and be coherent at both midsagittal and acoustic levels ensures the reliability of the area functions determined in such a way.

Zusammenfassung

Um die Kenntnisse der akustischen und artikulatorischen Zusammenhänge zu verbessern ist eine grosse Menge an Daten notwendig. Diese Arbeit verfolgt daher zwei Ziele: (1) Die Bestimmung von kohärenten Mittelsagittalfunktionen und den Querschnittsflächenfunktionen sowie den Formanten für eine geringe Anzahl von Vokalen und frikativen Konsonanten, bezogen auf eine Versuchsperson. (2) Es gilt, für die gleiche Versuchsperson, ein optimales Modell zur Transformation der Mittelsagittalfunktion in die Querschnittsflächenfunktion zu finden. Das verwendete Modell ist eine Erweiterung des $A = \alpha d^\beta$ Querschnittsflächenfunktion Modells von Heinz und Stevens, wobei α ständig vom mittelsagittalen Abstand sowie seiner Position von der Vokaltraktmittellinie abhängt. Die Koeffizienten

^{*} This paper is based on a communication presented at the ESCA conference EUROSPEECH-93 and has been recommended by the EUROSPEECH-93 scientific program committee.

^{*} Corresponding author. Fax: 33-76574710; E-mail: beautemps@icp.grenet.fr.

des Modells werden durch einen Optimierungsalgorithmus unter Verwendung der Methode der Gradientenminimierung ermittelt. Der Gradient des Fehlers zwischen den gewünschten Formanten und den tatsächlichen Formanten wird mit Hilfe eines Backpropagationnetzwerkes bestimmt. Das Netzwerk berechnet die Transformation der Mittelsagittalfunktion in die Querschnittsflächenfunktion sowie die Ausbreitung der Schallwellen. Die Tatsache, daß dieses Modell für die unterschiedlichen Konfigurationen Vokale oder frikative Konsonanten kohärent ist und die Modellbildung unter Berücksichtigung der Mittelsagittalfunktionen sowie der akustischen Parameter erfolgt, stellt eine Querschnittsflächenfunktion sicher, die einer guten Näherung der realen Situation entspricht.

Résumé

Dans le but d'améliorer les connaissances sur les relations articulatoire-acoustiques, d'importantes quantités de données sont nécessaires. L'objectif de ce travail a donc été double: (1) obtenir pour un même sujet, un ensemble cohérent de fonctions sagittales, fonctions d'aire et formants sur un nombre réduit de voyelles et consonnes fricatives, et (2) déterminer un modèle de passage de la fonction sagittale à la fonction d'aire valable pour ce même sujet. Des vues radiographiques du conduit vocal et le son associé sont disponibles pour ce locuteur ainsi que des vues de face des lèvres et des moulages du palais dur. Le modèle utilisé est une extension du modèle de fonction d'aire $A = \alpha d^\beta$ de Heinz et Stevens modifié pour que α dépende continûment de la position le long du conduit et de la distance sagittale. Les coefficients du modèle ont été déterminés par un algorithme d'optimisation fondé sur une technique de descente de gradient. Le gradient de l'erreur entre les formants désirés et les formants obtenus est déterminé à l'aide d'un réseau à rétro-propagation incluant le passage de la fonction sagittale à la fonction d'aire ainsi que la propagation des ondes acoustiques. Le fait que le modèle soit déterminé pour des configurations aussi différentes que des voyelles et des consonnes et qu'il soit cohérent aussi bien sur le plan sagittal qu'acoustique devrait assurer que les fonctions d'aire obtenues soient relativement proches de la réalité.

Keywords: Articulatory-acoustic relationship; Midsagittal profile; Area function; Articulatory data

1. Introduction

In the domain of speech production research, a good knowledge of the articulatory-acoustic relationships is a key factor. A number of studies deal with articulatory concepts and data, i.e. the positions of different articulators that finally determine the geometrical shape of the vocal tract, on the one hand, and with acoustic concepts and data, i.e. mainly the resonance frequencies of the vocal tract, on the other hand. In our quest towards understanding and modelling the articulatory-acoustic relationships, we are faced with the problem of deriving area functions from midsagittal profiles, these latter being obtained either from articulatory models or from teleradiographs.

The aim of this paper is to describe our attempts to establish a better model of midsagittal profile to area function conversion, and to derive reliable area functions for a set of sustained

vowels and fricative consonants uttered by a subject.

Two different approaches for deriving area functions for the production of speech sounds by subjects emerge from the literature: direct methods involving geometrical measurements of the vocal tract, and indirect methods based on acoustic inversion.

Among the direct measurement methods, only one can provide a complete three-dimensional picture of the vocal tract: Magnetic Resonance Imaging (cf. (Baer et al., 1991) for a careful survey), X-ray tomography of the entire vocal tract being considered too potentially hazardous to be of any practical use. The other direct methods do provide partial information only. X-ray radiography is widely spread; it delivers midsagittal profiles of the vocal tract (see e.g. (Fant, 1960; Bothorel et al., 1986)) and can be associated with optical lip recordings (Bothorel et al., 1986; Badin,

1991); computer-aided X-ray tomography can also provide very useful data (Sundberg, 1969; Sundberg et al., 1987; Perrier et al., 1992). Palatography or electropalatography furnishes essentially the contours of tongue contact with the hard palate (Hardcastle et al., 1991), and is of limited interest for the reconstruction of the vocal tract shape, unless associated with ultrasound imaging techniques (see e.g. (Stone et al., 1991)). Ultrasound techniques supply information on the tongue surface shape (Stone et al., 1988), but only in restricted zones of the tongue. The realisation of casts of the vocal tract of the subject constitutes a very informative method, however limited by its invasive character (Sundberg and Lindblom, 1990). Finally, optical measurements of the lips particularly provide a few important parameters, such as intralabial area or height, width and depth of the labial horn (Abry and Boë, 1986; Lallouache, 1990; Carter et al., 1990).

Several authors have studied the conversion from midsagittal profiles to area functions (Heinz and Stevens, 1965; Sundberg, 1969; Maeda, 1972; Sundberg et al., 1987; Baer et al., 1991; Fant, 1992; Perrier et al., 1992), but very few have thoroughly verified the accuracy of the formants computed from the area functions in relation to the formants measured simultaneously with the midsagittal profile (Sundberg, 1969; Maeda, 1972; Baer et al., 1991). Moreover, only vowels have been studied.

Indirect methods consist in determining the area function from acoustic data, either from the speech signal or from the acoustic response of the vocal tract to an external excitation (cf. e.g. (Schroeder, 1967; Sondhi and Gopinath, 1971; Sondhi and Resnick, 1983)). The resulting area functions will be suited to generate the exact formants through an acoustic model, by the very fact that they have been determined from these formants, but nothing ensures that they are the true area functions.

Our ambition has been to combine both approaches, and to derive realistic area functions coherent with measured midsagittal profiles and producing, through acoustic models, the formants measured from the sounds recorded simultaneously with the radiographs. Therefore, we have

obtained, for a French male subject articulating a set of sustained vowels and voiceless fricatives, simultaneous telerradiographs and sound recordings. From these initial data, midsagittal profiles and formant frequencies have been determined.

2. Derivation of midsagittal profiles and midsagittal functions

2.1. Teleradiography

In order to obtain midsagittal profiles for the different sustained configurations, conventional X-ray pictures of the vocal tract have been obtained for the subject during the sustained production of the vowels [a, i, u] and of the voiceless fricatives [f, θ, s, ʃ, ç, x]. The method used was teleradiography. The source was 5 m from the subject and the film immediately behind the subject's head, ensuring a negligible optic distortion; an aluminium filter was used in order to obtain a good contrast on soft parts such as the lips. A frontal photograph of the lips was taken immediately after each X-ray and simultaneous audio recordings were made (see (Badin, 1991) for more details).

2.2. Determination of midsagittal profiles

Up to now, there have been no automatic and reliable methods to determine the midsagittal profiles of the vocal tract from conventional X-ray pictures. These profiles have thus been traced by hand on transparent mylar sheets, as well as the inner contour of the lips on the front photograph. Due to the aluminium filter, the lip outline was easy to recover. Dental impressions of the upper teeth, with hard palate, and lower teeth with floor of the mouth, have been taken in order to improve the accuracy in the corresponding regions of the vocal tract.

The manually traced midsagittal profiles are then digitised into a list of contiguous pixels coded by their x - y coordinates. This operation is performed by a system developed at the Laboratoire de Traitement de l'Image et de Reconnaissance des Formes de l'Institut National Polytech-

nique, Grenoble, by S. Olympieff, involving a video camera and an automatic contour detection software.

2.3. Determination of midsagittal functions

2.3.1. Coordinate system

The midsagittal function is defined as the distance between the upper and lower contours of the midsagittal profile, measured on the line perpendicular to the vocal tract midline, as a function of the abscissa on this midline. In order to obtain such a function, the midsagittal profile is decomposed in a number of sections for which length and coronal height are determined. The coordinate system initially proposed by Heinz and Stevens (1965), and widely used (e.g. (Lindblom and Sundberg, 1971; Maeda, 1989; Perrier et al., 1992)) has been employed. This system is based on a grid that divides the vocal tract into sectors; each section of the midsagittal profile then corresponds to the vocal tract zone comprised between the two straight lines which define the sector. The semi-polar grid system is made up of three different parts: (1) a part located between the glottis and the low pharynx, made of parallel lines; (2) a part located between the low pharynx and the middle of the mouth cavity, made of straight lines converging in a single point, that serves as the origin of the coordinate system; and (3) a set of parallel lines between the middle of the mouth cavity and the end of the lips. This can be seen in Fig. 1.

The grid system needs to be laid, for each picture, in the same position in relation to fixed anatomical landmarks related to the subject's cranium and spinal column. A vertical axis parallel to the back pharynx wall has been chosen. The centre of the coordinate system and the angle that marks the limits of the polar coordinates zone have been determined so that this centre coincides with the centre of the circle that fits at best the velum and the hard palate contours, and so that the axis of the mouth linear coordinates zone is parallel to the mouth midline.

The different sections are then determined as the closed contours delimited by the segments belonging to the tract upper and lower contours,

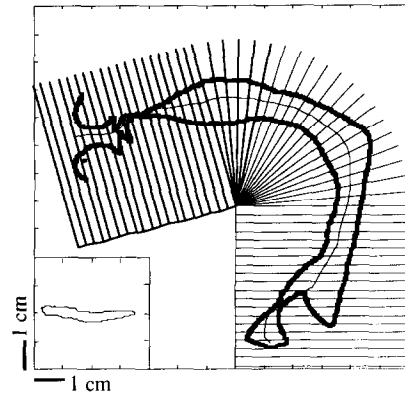
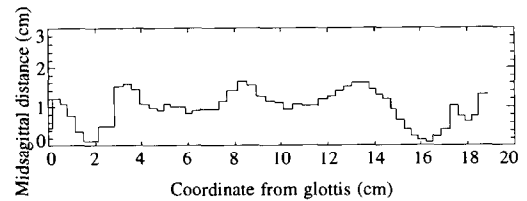


Fig. 1. Midsagittal profile with grid and centres of gravity overlaid (middle), lip inner contour (bottom left) and resulting midsagittal function for [s] (top).

and by the lines of the grid. For the section closest to the lips, the line tangent to both lips is used to close the contour.

2.3.2. Vocal tract midline and length

The determination of the midline of the vocal tract is not a trivial problem. Most vocal tract acoustic models take into account the propagation of plane waves only. Thus, the midline would ideally be the line for which the wave front at any point is always perpendicular to the tangent to the line at this point. For a tract with limited discontinuities, the midline is not difficult to determine. However, there are some discontinuities in real vocal tracts: at the level of the piriformis sinuses the laryngeal vestibule opens into a much larger zone, and the front cavity of a [ʃ] including the sublingual cavity presents also a large discontinuity. A first, relatively well known, problem is that of the internal length correction (see (Sundberg et al., 1992)). When the cross-sectional area of a tube is small enough compared to the cross-sectional area of the tube it is connected with, an

Table 1
Vocal tract lengths of the nine configurations of the corpus

Config.	Initial estim.	Final estim.	TB(x)	TB(MRI)	PN(x)	PN (MRI)
[a]	19.14	18.06	17.6	17.5	16.8	15.75
[i]	18.84	17.50	18.1	16.62	17.7	16.62
[u]	20.33	18.75	18.3	18.37	18.7	18.37
[f]	18.70	17.51				
[θ]	18.73	17.32				
[s]	18.85	17.65				
[ʃ]	20.03	18.18				
[ç]	18.61	17.21				
[x]	18.92	17.65				

First column: initial estimation from the center of gravity algorithm; second column: final estimation after corrections at the lips and at the larynx. The other columns correspond to the data published by Baer et al. (1991, p. 813).

internal end correction should be applied to the small area tube, according to the formula derived by Ingård (1953), quoted by Fant (1960):

$$l_c = 0.48\sqrt{A_0} \left[1 - 1.25\sqrt{\frac{A_0}{A}} \right], \quad (1)$$

provided that the aperture area A_0 is smaller than $0.16A$. This formula is valid if the two tubes have the same axis; if the small area tube is strongly eccentric, an extra lengthening should be taken into account (cf. the experiment made by Holmes (1981)).

For each of the sections, the surface and centre of gravity of the corresponding closed contours is evaluated by a pixel counting algorithm. The vocal tract midline is then drawn as the line joining the centres of gravity. The length of each section is computed as the sum of the lengths of the two straight segments of the midline located within the section. Finally, the midsagittal distance is calculated as the ratio of the surface of the section over its length. An evaluation of this *centre of gravity algorithm* is reported in (Badin and Tzavali, 1990): the length of each section appears to stabilise for distances between two neighbouring lines of the grid smaller than 4 mm, and a distance of 3 mm is a good compromise between the accuracy needed for the sampling of the area function (Wu et al., 1987) and the noise induced by the finite dimension of the pixels. Fig. 1 shows an example of result for the configuration [s].

The initial estimations of the vocal tract lengths obtained by this method are given for the nine configurations of the corpus in Table 1. The last lip section has finally been removed to build the area functions, since it would correspond to an area too wide to be representative of a radiating area. The larynx region had to be represented by a single 2 cm long tube (see below); this has contributed to shorten somehow the vocal tract, since the original larynx length was slightly longer (around 2.5 cm). The final estimates of this length is also given in Table 1. In two cases, i.e. for [u] and [ʃ], one of the sections has been manually shortened from about 6 mm to about 3 mm, in order to reduce the possibly exaggerated effect of the sudden jump in the midsagittal profile in the sublingual cavity region. Finally, the teeth apparent in the midsagittal profile of [ʃ] have been removed, since their acoustical effects are not really known, and likely small.

3. Previous α , β models of midsagittal to area function conversion – available data

A review of the literature in the domain of midsagittal function to area function conversion models shows that all the solutions evolve around the model that will be referred to as the α , β model, proposed for the first time by Heinz and Stevens (1965). This model relates the cross-sectional area A at any given point of the vocal tract

along its midline to the midsagittal distance d measured at that point by $A = \alpha d^\beta$, where α and β are coefficients depending on the subject and on the location in the vocal tract. This model has been used and refined by a number of researchers (Sundberg, 1969; Maeda, 1972; Sundberg et al., 1987; Baer et al., 1991; Perrier et al., 1992; Fant, 1992). The crucial problem is to determine the values of those α and β coefficients.

Sundberg (1969) used two sets of α and β coefficients, one for the mouth region, and the other for the pharynx region. The coefficients for the mouth have been determined for three male subjects by measurements on plaster casts of the mouth, assuming a flat tongue shape. The coefficient α was in the range (2.07–2.63), and β in the range (1.33–1.47). For the pharynx region, the cross-sectional areas were taken as the area of an ellipse, of which one axis was the midsagittal distance and the other one varied between 3 cm in the upper pharynx down to 1.5 cm in the larynx. Sundberg (1969) also commented earlier data obtained by Fant (1964): “the side walls of the pharynx cavity tend to approach each other [...], when the distance between the tongue and the pharynx wall exceeds a certain value (1.9 cm)”. This statement was disproved by Sundberg et al. (1987, p.83). With this two-region model, the computed formants for the nine long Swedish vowels deviated from the measured ones by less than 15% for F_1 , 5% for F_2 and 7% for F_3 .

Sundberg et al. (1987) have extended Sundberg’s (1969) work to computer radio-tomographies of the pharyngeal region. They also derive α and β coefficients. Fant (1992) also derived α and β values which turned out similar to those of Sundberg. Fit of calculated values versus measured was also discussed.

Baer et al. (1991) have attempted to establish α coefficients assuming a constant $\beta = 2$, and also α and β coefficients, for each individual 1/2 cm slice of their 3D magnetic resonance imaging data. Their formants are however not in very good agreement with the values measured on the same subject.

A comparison of these data, whenever it has been possible, is presented in the discussion in the last part of this paper.

Maeda’s approach (1972) is relatively different. Quoting Chiba and Kajiyama (1941) and Fant (1960), he suggested that information only based on midsagittal distance at any given point does not specify the cross-sectional area completely. He first established, by means of an optimisation procedure, a set of α coefficients with a bi-parabolic assumption (leading to $\beta = 1.5$) that would provide the best fit with the measured formants. In a second step, another optimisation procedure led to the determination of the a_n , b_n , c_n and k_n coefficients of the relation giving the transversal distance s_n as a linear function of the midsagittal distance at the same location, d_n , and also of two other midsagittal distances, d_p and d_q , supposed to represent the influence of the phonetic quality of the vowel:

$$s_n = a_n d_n + b_n d_p + c_n d_q + k_n. \quad (2)$$

Finally, the area is computed as $A_n = (\pi/4)d_n s_n$. The values for the first three formants were in a good agreement with the measured values. There was, however, no confrontation with real geometrical data.

In their last investigation on that problem, Perrier et al. (1992), taking into account Maeda’s statement, proposed a model inspired by the α , β model. Based on analysis of CT-scans obtained for the three cardinal vowels [i, a, u] pronounced by a male subject, they determined the coefficient α as a function of the vocal tract region where the midsagittal distance was measured, with the aim to globally integrate the phonetic nature of the vowel. For each of the seven regions into which they divided the vocal tract, the coefficient α is computed as a linear interpolation between two extreme values α_{inf} and α_{sup} , related respectively to small and large midsagittal dimensions and determined for that given region. Values of the immediately adjacent sections are also taken into account (see (Perrier et al., 1992) for more details). The coefficient β was fixed to 1.5 in reference to the bi-parabolic geometric model of the shape of the vocal tract (Maeda, 1972; Perrier et al., 1992). This model led for vowels to a good prediction of area functions, both from acoustical and articulatory points of view.

4. A new model: extension of Perrier et al.'s region model

The present work has partly been motivated by the fact that we did not succeed in extending Perrier et al.'s model (1992) to front fricatives. As stated in the introduction, our problem was to coherently derive a set of area functions from the midsagittal functions and formants simultaneously measured on a subject, for a corpus of vowels and voiceless fricative consonants. Preliminary experimentation has convinced us that the extension to consonants of the Perrier et al.'s model – designed for vowels – was rather tricky. More precisely, Perrier et al.'s model gave fairly good first formants for vowels [a, i] (error less than 10 Hz), as well as for fricatives [x, ç] – which have articulations close to those of [a, i] (error less than 20 Hz) – but rather large errors (about 90 Hz) for the front fricatives [f, θ, s, ʃ]. Large errors have been also noticed on the other formants (200 Hz on average for F_2 and F_3 , for instance) for all the configurations. A new extension of this model has thus been developed.

The main difference between the proposed model and Perrier et al.'s is that the α_{inf} and α_{sup} coefficients are not associated to vocal tract regions, but continuously dependent of the corresponding coordinate along the vocal tract midline. This choice saves us from having to determine manually, and somehow arbitrarily, the limits between the different zones, and from encountering discontinuities between regions. Generally speaking, any arbitrary function defined over a limited domain of x ($0 \leq x \leq l_{\text{tot}}$) can be expressed as a Fourier series. The α_{inf} and α_{sup} coefficients have thus been defined as follows:

$$\alpha_{\text{inf}}(x) = a_{\text{inf},0} + \sum_{n=1}^3 [a_{\text{inf},n} \cos(n\omega x) + b_{\text{inf},n} \sin(n\omega x)] \quad (3)$$

and

$$\alpha_{\text{sup}}(x) = a_{\text{sup},0} + \sum_{n=1}^3 [a_{\text{sup},n} \cos(n\omega x) + b_{\text{sup},n} \sin(n\omega x)], \quad (4)$$

where $\omega = \pi/l_{\text{tot}}$ and l_{tot} is the total length of the vocal tract. The choice of the number of harmonics has been dictated by a trade-off between accuracy of approximation and number of parameters to infer. The amount of data to represent being 45 in our case (9 configurations times 5 formants), the use of three harmonics (leading to 14 independent parameters) has been deemed optimal.

The coefficient $\alpha(d, x)$ is then computed as a linear interpolation between $\alpha_{\text{inf}}(x)$ and $\alpha_{\text{sup}}(x)$ between the midsagittal distance thresholds d_{inf} and d_{sup} :

$$\alpha(d, x) = \alpha_{\text{inf}}(x) \quad \text{if } d < d_{\text{inf}}, \quad (5)$$

$$\alpha(d, x) = \alpha_{\text{inf}}(x) + \frac{\alpha_{\text{sup}}(x) - \alpha_{\text{inf}}(x)}{d_{\text{sup}} - d_{\text{inf}}} (d - d_{\text{inf}}) \quad \text{if } d_{\text{inf}} \leq d \leq d_{\text{sup}}, \quad (6)$$

$$\alpha(d, x) = \alpha_{\text{sup}}(x) \quad \text{if } d > d_{\text{sup}}. \quad (7)$$

Perrier et al. (1992) found that reasonable values for d_{inf} and d_{sup} are respectively 1 and 2 cm. We have thus used the same thresholds in the new model. Finally, $A(x)$ is simply computed as

$$A(x) = \alpha(d, x) d(x)^\beta \quad \text{with } \beta = \text{const.} \quad (8)$$

It should be mentioned that the larynx region was not included in the model, because its pyramid-shaped structure could not be correctly mapped by the α, β model, and that it was thus represented by a fixed uniform tube of 1.8 cm² area and 2 cm length, following Fant's suggestion (1960). Oppositely, the lip region was included in the model.

5. Experimental data and optimisation of the model parameters

Once the model is defined, the problem one is faced with is to optimise its parameters for a given subject. The model is actually entirely defined by the two sets of seven Fourier series coefficients controlling the $\alpha_{\text{inf}}(x)$ and $\alpha_{\text{sup}}(x)$ curves. These coefficients must be optimised for the whole corpus, i.e. they should be chosen such

as the four or five first formants computed from the sagittal functions through the new sagittal-to-area conversion model and the acoustic model fit at best the formant frequencies measured on the subject, for each of the nine configurations. This optimisation will lead indirectly to the determination of the area functions, that can be viewed as intermediate parameters between the sagittal-to-area model and the acoustic model. Note that if these area functions had to be determined directly from formants, we would be confronted to the well established fact that the articulatory-to-acoustic conversion is a many-to-one function (see e.g. (Atal et al., 1978)), and that the inversion of such a function is an *ill-posed problem*. A problem is *ill-posed* in the mathematical sense if it is not certain that the solution exists, is unique, and continuously dependent on the initial conditions (Lavrentiev, 1967). The approach proposed here both provides the articulatory-like constraints needed to regularise the problem and to find solutions which are compatible with a human vocal tract, and establishes a model optimised for the subject.

Having in mind that the 50 section lengths $l_{k,i}$ and midsagittal distances $d_{k,i}$ are given and fixed, as well as the four or five first formant frequencies $F_{k,j}$, the fact that the $\alpha(d,x)$ coefficients are not independent of each other, but linked by a relatively smooth Fourier series function, clearly constitutes a strong constraint, since the 50 section areas $A_i(x)$ will be entirely determined by only two sets of seven Fourier coefficients. Another constraint is that the Fourier series coefficients are used not for a single configuration, but for the whole set: the model is therefore more likely to work for the vowels *and* for the consonants of the corpus.

5.1. Geometrical data

Other constraints can be provided by the knowledge of some experimentally measured geometrical parameters. For the same subject and the same sounds, aerodynamic equivalents of the constriction area have thus been determined from flow and pressure measurement in the vocal tract for [f, θ, s, ʃ, ç] (see (Castelli et al., 1990)). Lip

areas have also been measured simultaneously with the midsagittal profiles (Badin, 1991). The deviation of the actual values obtained for the area function from these target values have been used as an additional error function for the optimisation algorithm.

5.2. Formants frequencies

According to the theory of fricative production (see e.g. (Badin, 1989, 1991)), the transconductance between the acoustic flow at the lips and the noise pressure source created somewhere inside the vocal tract will exhibit poles and zeros. However, if the constriction area is small enough (which is the case for the fricatives under study), the poles affiliated with the cavity behind the constriction will be nearly cancelled by *bound* zeros associated with the same back cavity. Thus, only the poles associated with the front cavity will remain visible in this transconductance, and in the spectrum of the voiceless fricatives. Since the formants of the whole vocal tract were needed for the optimisation process, these formants have been measured as the poles of the transfer function between the acoustic flow at the lips and that at the glottis, by means of a direct measurement technique (Badin, 1991). The results are given in Table 2.

5.3. Network implementation of the direct models

Once the constraints are made explicit, the problem reduces to the optimisation of two sets of seven Fourier series coefficients. The optimal set of coefficients will minimise the quadratic error between actual and desired formants for all configurations in the data. The total chain of transformations in the model is shown in Fig. 2(a). The various functional components can be distinguished: estimation of $\alpha_{k,\text{inf}}$ and $\alpha_{k,\text{sup}}$ through Fourier series, determination of α_k by interpolation, computation of $A_k(x)$ with the α , β equation, and evaluation of the error between the desired frequencies $F_{k,j}^*$, and the attained frequencies $\hat{F}_{k,j}$ where k ($1 \leq k \leq 9$) is the configuration and j ($1 \leq j \leq 5$) the formant index.

Table 2
Measured target formants (first column) and formants obtained after midsagittal distance adjustment (second column); the symbol • indicates that no value could be measured

Configurations	Target formants (Hz)	Reached formants (Hz)
[a]	<i>F</i> 1 590 <i>F</i> 2 1260 <i>F</i> 3 2415 <i>F</i> 4 3533 <i>F</i> 5 • <i>F</i> 6 5300	528 1305 2388 3544 4055 5295
[i]	<i>F</i> 1 253 <i>F</i> 2 2077 <i>F</i> 3 3100 <i>F</i> 4 3800	297 2072 3062 3708
[u]	<i>F</i> 1 326 <i>F</i> 2 812 <i>F</i> 3 2320 <i>F</i> 4 3450	367 877 2325 3472
[f]	<i>F</i> 1 397 <i>F</i> 2 1282 <i>F</i> 3 2504 <i>F</i> 4 3665	413 1392 2463 3662
[θ]	<i>F</i> 1 386 <i>F</i> 2 1456 <i>F</i> 3 2677 <i>F</i> 4 3869	346 1476 2665 3880
[s]	<i>F</i> 1 400 <i>F</i> 2 1500 <i>F</i> 3 2647 <i>F</i> 4 • <i>F</i> 5 4276 <i>F</i> 6 5062	373 1452 2630 3902 4222 5046
[ʃ]	<i>F</i> 1 450 <i>F</i> 2 1710 <i>F</i> 3 2230 <i>F</i> 4 2952	421 1703 2315 2959
[ç]	<i>F</i> 1 305 <i>F</i> 2 2036 <i>F</i> 3 3015 <i>F</i> 4 3730	310 2069 2981 3632
[x]	<i>F</i> 1 600 <i>F</i> 2 1211 <i>F</i> 3 2180 <i>F</i> 4 3665	541 1227 2176 3663

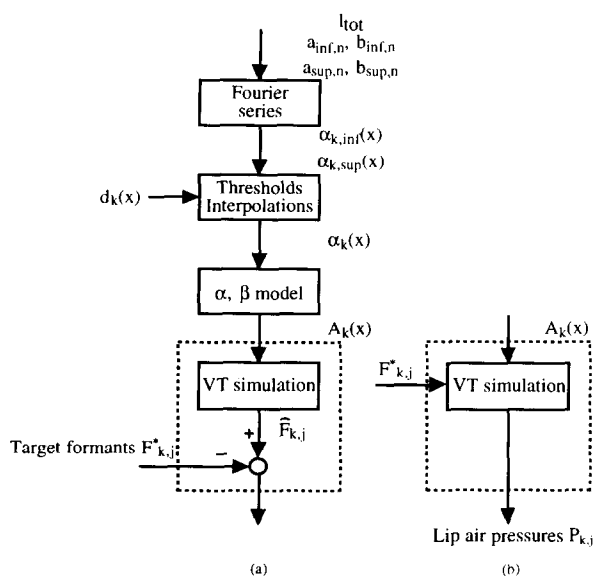


Fig. 2. Graphs implementing the midsagittal-to-area and acoustic wave propagation models.

For a given set of data corresponding to configuration indexed $k \langle d_{k,1}, \dots, d_{k,N}; l_{k,1}, \dots, l_{k,N}; F_{k,1}, \dots, F_{k,5} \rangle$, several optimisation algorithms can be used to minimise the error:

$$E = \sum_{k=1}^9 e_k^2, \tag{9}$$

with respect to the Fourier series coefficients (e_k is the contribution of the configuration k to the global error E). Evidently, the gradient of the error cannot be easily derived because the algorithm implementing the area-to-formants model cannot be described in an analytical way and involves a Newton-Raphson-type iterative technique. In order to overcome this difficulty, and to benefit from the convergence properties of gradient descent-based algorithms, a different way to implement the acoustic model has been adopted. The amplitude of the pressure along the vocal tract is computed recursively from a unit amplitude excitation at the glottis, using the fact that the ratio of the pressure derivatives on each side of the junction between two sections is inversely proportional to the ratio of the areas. A lossless

case is assumed. The radiation phenomenon is taken into account as an end correction:

$$l_c = 0.8\sqrt{A_0/\pi}, \quad (10)$$

where A_0 is the lip area. The finite wall impedance is represented by a frequency correction:

$$F_{jw} = \sqrt{F_{j0}^2 + F_{w0}^2}, \quad (11)$$

where F_{jw} is the corrected frequency, F_{j0} is the hard-walled frequency and F_{w0} , fixed to 190 Hz, is the closed lip resonance frequency (Badin and Fant, 1984). As it can be shown that the formants are the frequencies for which a pressure node occurs at the lip end of the vocal tract, the quadratic error of the pressure at the lips can be used in place of the quadratic error between desired and attained formants. The new acoustic model then computes the pressure at the lips $P_{k,j}$ from the area functions and the desired formant frequencies $F_{k,j}^*$ (see Fig. 2(b)). The advantage is that the computation of the lip pressure is entirely analytical. Hence, it is possible, to determine the gradient of the error:

$$E = \sum_{k=1}^9 e_k^2, \quad \text{where } e_k^2 = \sum_{j=1}^5 P_{k,j}^2, \quad (12)$$

with respect to the area of each section, and thus to each Fourier transform coefficient.

Note that the error E obviously depends on the pressure distribution in the vocal tract, which is related to each configuration. It is thus difficult to determine explicitly the formant-frequency weighting of this error.

Instead of deriving analytically the gradient, a technique reminiscent of the error back-propagation method, popular in the neural nets research community (see e.g. (Rumelhart et al., 1986)) was used. According to this technique, the computation of the pressure amplitude from section to section can be represented as a graph of nodes which compute analytical functions of their inputs. From such a graph, it is possible to derive another one, called *dual graph*, in which each node is replaced by the equivalent derivatives of

the output of the node with respect to its inputs, and the flow of signal is done backwards. In essence, if the error value is fed in the input of the dual graph (i.e., the equivalent output of the original graph), the back-propagation operations will yield the derivative of this error with respect to each arbitrary points of the graph. Besides the evident computational advantage, this approach offers high flexibility, because the changes in the original graphs can be automatically taken into account in the dual graph for the computation of the gradient. This flexibility is highly desired because, as it will be noted below, the optimisation procedure must be finely controlled in order to obtain satisfactory results. Further details concerning this algorithms can be found in (Laboissière, 1992).

The total error function integrates the lip pressure error and the other explicit constraints: the cost increases when the constriction area is outside a given range (set by A_{cmin} and A_{cmax} , see further Fig. 6(a)), and when the lip area deviates from the lip target area (see Fig. 6(b)).

5.4. Optimisation phases

The optimisation process has been achieved in two phases. The first and most important phase aimed at deriving the model parameters, whereas the second phase was devoted to minor corrections of the original midsagittal functions.

5.4.1. Model parameters optimisation

During the first phase, nine identical networks such as the one described in Fig. 2 were operated in parallel, in order to optimise the Fourier series coefficients over the complete corpus. The error back-propagated through the networks is the sum of the errors computed by each network for each configuration. All the constraints were applied, i.e. area constraints for the constriction region and lip region (see Fig. 6) including the aerodynamic equivalent of the constriction area available for [f, θ, s, j, ç], the given ranges in which area constrictions are expected to fall for all the configurations, and also lip target areas.

For this phase, as well as for the second one, β has been set to a value of 1.5, since preliminary

experiments had shown that this was a reasonable value (cf. (Perrier et al., 1992)). Note that the α and β parameters are somehow mutually dependent: a high value of α could be compensated by a low value of β (see further the discussion). The target formant frequencies, the midsagittal distances and the section lengths are fed into the network, and frozen. The algorithm optimises iteratively the Fourier series coefficients, which thus defined a set of $\alpha_{\text{inf}}(d, x)$ and $\alpha_{\text{sup}}(d, x)$ curves. However, the results were not entirely satisfactory: fairly good formant frequencies were obtained for [a, i, f, s, θ , ζ], but fairly important mismatches affected [u, \int , ζ].

As it had been noticed that the system was finally quite strongly constrained (there were only 14 free input parameters but 45 output formant frequencies, 9 configurations times 5 formants), an attempt had been made to increase the degrees of freedom by using more Fourier coefficients. Results were not improved by using more

Fourier coefficients. It had thus been conjectured that there may be some minor errors in the midsagittal functions, that could be partly attributed to the manual acquisition and interpretation of the vocal tract outline. Thus it was decided to pursue further the optimisation by slightly adjusting the midsagittal distances.

5.4.2. Adjustment of the midsagittal functions

During the second optimisation phase, the target formant frequencies, the section lengths and the Fourier series coefficients are fed into the network, and frozen, i.e. the model is fixed. In this case, the 50 midsagittal distances are the parameters to optimise, as in a classical inversion scheme. Each configuration is processed separately, the constraints on the geometry being removed. The optimisation process was allowed a few iteration steps only, and the midsagittal distances were thus slightly adjusted, but kept into physiologically plausible limits.

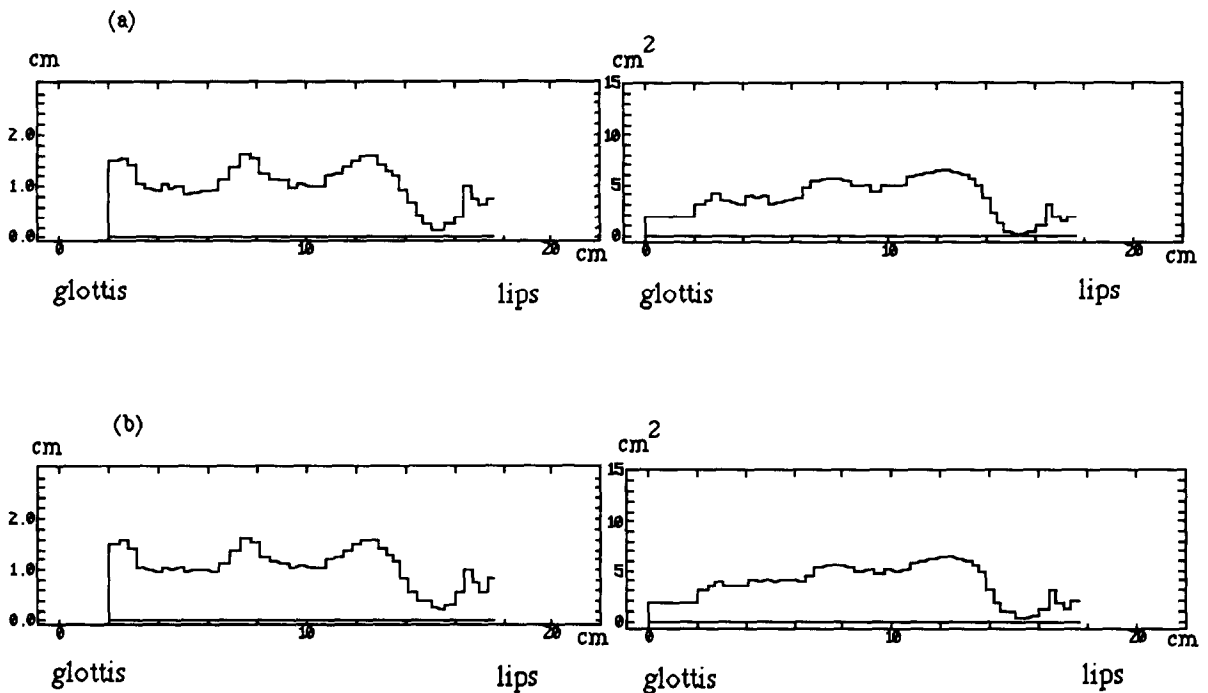


Fig. 3. Example of (a) original and (b) corrected midsagittal functions (left) and associated area functions (right) versus coordinate from the glottis, for [s].

Table 3
Midsagittal and area functions for the nine configurations of the corpus

[a]			[i]			[u]		
<i>d</i>	<i>A</i>	<i>X_c</i>	<i>d</i>	<i>A</i>	<i>X_c</i>	<i>d</i>	<i>A</i>	<i>X_c</i>
●	1.8	2	●	1.8	2	●	1.8	2
1.376	2.822	2.363	1.808	2.369	2.314	2.103	1.449	2.331
1.506	3.334	2.681	1.636	3.335	2.623	1.902	2.545	2.651
1.183	3.396	3.003	1.686	3.892	2.936	1.81	3.478	2.982
1.025	3.248	3.321	1.476	4.258	3.247	1.485	4.069	3.311
0.753	2.191	3.63	1.575	4.745	3.551	1.524	4.49	3.628
0.846	2.726	3.934	1.533	4.992	3.851	1.418	4.615	3.934
0.93	3.252	4.235	1.724	5.545	4.161	1.613	5.217	4.234
1.104	4.125	4.537	1.81	5.872	4.478	1.495	5.225	4.544
1.093	4.184	4.839	2.257	7.319	4.786	1.421	5.209	4.866
1.072	4.177	5.14	1.992	6.201	5.104	1.173	4.527	5.186
0.709	2.356	5.441	2.037	6.356	5.453	1.155	4.545	5.561
0.489	1.372	5.856	2.14	6.759	5.795	1.15	4.59	5.943
0.533	1.585	6.271	2.046	6.147	6.154	1.182	4.756	6.346
0.731	2.574	6.687	2.148	6.379	6.534	1.195	4.84	6.774
0.905	3.592	7.105	2.175	6.226	6.925	1.263	5.087	7.215
1.33	5.295	7.528	2.367	6.772	7.307	1.397	5.437	7.655
1.661	5.647	7.955	2.308	6.255	7.686	1.287	5.212	8.089
1.623	5.631	8.39	2.077	5.147	8.061	1.143	4.861	8.522
1.521	5.66	8.814	1.806	5.386	8.43	0.94	4.04	8.951
1.385	5.621	9.222	1.624	5.678	8.794	0.523	1.712	9.377
1.314	5.604	9.616	1.327	5.591	9.158	0.257	0.601	9.795
1.204	5.444	9.998	1.168	5.331	9.524	0.187	0.382	10.201
1.3	5.778	10.371	0.862	4.074	9.896	0.208	0.454	10.598
1.328	5.93	10.738	0.554	2.042	10.275	0.238	0.564	10.988
1.385	6.12	11.095	0.484	1.592	10.66	0.439	1.429	11.376
1.483	6.326	11.446	0.436	1.378	11.056	0.686	2.811	11.764
1.554	6.442	11.797	0.362	1.128	11.459	0.945	4.556	12.152
1.553	6.478	12.151	0.326	0.921	11.866	1.125	5.473	12.541
1.621	6.505	12.506	0.29	0.761	12.279	1.27	5.937	12.93
1.661	6.458	12.863	0.281	0.713	12.696	1.342	6.063	13.321
1.788	6.266	13.225	0.268	0.651	12.997	1.267	5.774	13.624
1.571	6.164	13.528	0.231	0.508	13.3	1.288	5.72	13.933
1.632	5.989	13.829	0.218	0.45	13.604	1.323	5.664	14.243
1.361	5.485	14.133	0.191	0.357	13.909	1.233	5.25	14.561
1.498	5.411	14.438	0.176	0.304	14.216	1.073	4.571	14.871
1.36	4.987	14.745	0.185	0.314	14.523	1.396	5.143	15.181
1.343	4.68	15.054	0.216	0.382	14.83	1.682	4.77	15.489
1.393	4.365	15.392	0.252	0.459	15.136	1.349	4.479	15.847
1.452	3.947	15.729	0.174	0.27	15.44	1.08	3.801	16.224
1.434	3.53	16.081	0.28	0.524	15.762	0.812	2.528	16.548
0.958	3.004	16.46	0.438	0.915	16.154	0.678	1.845	16.853
0.826	2.302	16.805	0.748	1.943	16.564	1.332	3.257	17.177
1.452	2.843	17.121	0.807	2.036	16.898	0.978	2.943	17.507
1.06	2.955	17.43	0.838	2.075	17.198	0.232	0.329	17.819
0.854	2.232	17.733	0.945	2.581	17.5	0.482	0.963	18.12
1.199	2.956	18.063				0.507	1.022	18.42
						0.495	0.977	18.751

6. Results and validation

In this section, the results are presented and analysed in the light of the previous studies and

of complementary data obtained from plaster casts of the subject. Figs. 3 and 4 are given as an illustration for the [s] configuration and for the three cardinal vowels [a, i, u].

Table 3 (Continued)

[f]			[θ]			[s]		
<i>d</i>	<i>A</i>	<i>X_c</i>	<i>d</i>	<i>A</i>	<i>X_c</i>	<i>d</i>	<i>A</i>	<i>X_c</i>
●	1.8	2	●	1.8	2	●	1.8	2
1.605	2.781	2.33	1.638	2.881	2.371	1.514	2.964	2.407
1.318	3.328	2.663	1.51	3.568	2.71	1.565	3.544	2.745
1.233	3.59	3.002	1.413	3.982	3.054	1.423	3.977	3.098
1.014	3.288	3.335	1.129	3.686	3.392	1.048	3.458	3.461
0.998	3.414	3.654	1.105	3.811	3.717	0.992	3.439	3.803
0.944	3.269	3.961	1.023	3.66	4.029	0.973	3.473	4.116
0.988	3.618	4.263	1.131	4.191	4.333	1.071	3.998	4.421
0.947	3.486	4.567	1.082	4.11	4.648	1.032	3.947	4.728
0.909	3.345	4.872	1.03	3.985	4.971	1.046	4.079	5.038
0.806	2.84	5.174	0.937	3.543	5.281	0.973	3.793	5.349
0.805	2.888	5.57	1.01	4.057	5.652	1	4.008	5.716
0.8	2.901	5.971	1.064	4.349	6.028	0.995	4.03	6.089
0.841	3.166	6.38	1.133	4.666	6.415	0.979	3.98	6.474
1.122	4.629	6.8	1.288	5.187	6.822	1.131	4.665	6.876
1.359	5.359	7.223	1.432	5.509	7.235	1.391	5.425	7.287
1.617	5.657	7.647	1.568	5.639	7.641	1.646	5.659	7.695
1.496	5.6	8.074	1.518	5.619	8.046	1.545	5.626	8.103
1.516	5.651	8.491	1.476	5.637	8.448	1.271	5.305	8.503
1.377	5.602	8.904	1.53	5.733	8.85	1.156	5.086	8.897
1.342	5.653	9.311	1.482	5.816	9.244	1.144	5.146	9.286
1.307	5.696	9.703	1.376	5.824	9.63	1.038	4.869	9.667
1.299	5.784	10.08	1.35	5.89	10.003	1.099	5.179	10.042
1.345	5.974	10.451	1.339	5.966	10.367	1.044	5.047	10.411
1.306	5.974	10.816	1.299	5.956	10.724	1.057	5.154	10.772
1.321	6.068	11.175	1.275	5.946	11.076	1.207	5.739	11.129
1.463	6.37	11.533	1.379	6.226	11.424	1.271	5.96	11.487
1.465	6.386	11.89	1.38	6.231	11.773	1.405	6.288	11.843
1.528	6.433	12.251	1.432	6.294	12.126	1.492	6.41	12.2
1.5	6.313	12.616	1.468	6.266	12.483	1.592	6.443	12.559
1.573	6.233	12.985	1.421	6.038	12.844	1.597	6.325	12.925
1.405	5.843	13.289	1.317	5.652	13.147	1.418	5.964	13.227
1.346	5.531	13.589	1.166	5.05	13.449	1.296	5.525	13.528
1.304	5.224	13.89	0.983	4.202	13.75	1.159	4.954	13.83
1.148	4.619	14.191	0.82	3.087	14.051	0.822	3.158	14.132
1.013	4.022	14.493	0.721	2.436	14.352	0.566	1.737	14.435
0.905	3.281	14.797	0.569	1.627	14.653	0.407	1.016	14.738
0.803	2.621	15.101	0.487	1.216	14.953	0.382	0.882	15.039
0.789	2.424	15.43	0.416	0.887	15.253	0.245	0.433	15.34
1.228	3.502	15.759	0.403	0.794	15.555	0.237	0.394	15.646
1.3	3.27	16.077	0.254	0.403	15.859	0.306	0.548	16.023
0.341	0.594	16.498	0.191	0.25	16.278	0.546	1.238	16.414
0.338	0.566	16.89	0.324	0.531	16.698	1.02	2.987	16.724
0.338	0.557	17.197	0.653	1.518	17.009	0.767	1.935	17.031
0.876	2.301	17.502	0.8	1.996	17.312	0.559	1.182	17.341
						0.833	2.134	17.646

6.1. Midsagittal functions

Fig. 3 illustrates the minor changes brought about to the midsagittal function in order to

increase the fit between the measured and attained formants. It can be seen that the changes have been made in the vicinity of the constriction, where the relative accuracy of hand tracings is

Table 3 (Continued)

[ʃ]			[ç]			[x]		
<i>d</i>	<i>A</i>	<i>X_c</i>	<i>d</i>	<i>A</i>	<i>X_c</i>	<i>d</i>	<i>A</i>	<i>X_c</i>
●	1.8	2	●	1.8	2	●	1.8	2
2.008	1.622	2.351	1.682	2.724	2.313	1.575	2.798	2.342
1.838	2.963	2.678	1.783	3.274	2.627	1.346	3.349	2.679
1.751	3.785	3.012	1.565	3.983	2.942	1.218	3.557	3.025
1.394	4.118	3.348	1.618	4.479	3.255	0.913	2.837	3.365
1.363	4.385	3.675	1.457	4.644	3.561	0.863	2.748	3.69
1.151	4.048	3.987	1.601	5.178	3.862	0.727	2.215	3.999
1.302	4.673	4.29	1.563	5.36	4.162	0.898	3.138	4.3
1.26	4.693	4.59	1.784	5.887	4.463	0.862	3.033	4.625
1.514	5.521	4.891	1.631	5.817	4.773	0.599	1.801	4.97
1.385	5.283	5.199	1.631	5.817	4.773	0.295	0.633	5.296
1.478	5.604	5.546	1.721	6.053	5.132	0.413	1.068	5.681
1.471	5.622	5.898	1.778	6.134	5.487	0.59	1.844	6.068
1.612	5.895	6.267	1.841	6.11	5.856	0.856	3.26	6.465
1.588	5.815	6.665	1.874	5.975	6.241	1.027	4.276	6.879
1.72	5.831	7.091	2.1	6.053	6.63	1.199	4.936	7.305
1.861	5.599	7.513	1.88	5.615	7.014	1.261	5.158	7.723
1.861	5.446	7.917	1.791	5.587	7.396	1.176	4.987	8.141
1.744	5.541	8.317	1.662	5.625	7.776	1.077	4.751	8.554
1.397	5.529	8.714	1.647	5.62	8.155	0.833	3.468	8.963
1.275	5.383	9.11	1.457	5.646	8.536	0.738	2.948	9.361
1.123	5.054	9.503	1.219	5.323	8.914	0.687	2.696	9.748
0.873	3.794	9.893	0.962	4.422	9.291	0.745	3.088	10.128
0.638	2.409	10.28	0.804	3.44	9.672	0.802	3.497	10.499
0.451	1.462	10.665	0.589	2.192	10.055	0.916	4.312	10.86
0.413	1.309	11.049	0.5	1.736	10.446	1.223	5.796	11.217
0.485	1.686	11.433	0.438	1.435	10.846	1.313	6.076	11.574
0.607	2.361	11.821	0.369	1.112	11.252	1.411	6.297	11.93
0.653	2.621	12.213	0.342	0.988	11.662	1.547	6.459	12.29
0.755	3.224	12.61	0.306	0.825	12.076	1.568	6.403	12.651
0.816	3.561	13.012	0.343	0.959	12.491	1.617	6.291	13.019
0.644	2.457	13.312	0.253	0.595	12.792	1.578	6.101	13.324
0.564	1.973	13.612	0.279	0.672	13.096	1.614	5.893	13.625
0.401	1.145	13.912	0.25	0.553	13.403	1.375	5.417	13.929
0.366	0.941	14.213	0.186	0.183	13.714	1.481	5.27	14.234
0.391	1.039	14.522	0.179	1.065	14.028	1.323	4.81	14.541
0.299	0.665	14.872	0.133	0.19	14.344	1.375	4.562	14.847
0.989	3.825	15.222	0.182	0.291	14.665	1.363	4.227	15.156
1.413	4.277	15.572	0.216	0.357	15.016	1.284	3.885	15.472
1.203	3.849	15.922	0.445	0.996	15.363	1.552	3.139	15.796
1.128	3.535	16.258	0.859	2.335	15.676	1.416	3.142	16.116
1.128	3.333	16.603	0.624	1.513	15.988	0.921	2.697	16.423
1.495	2.815	16.947	0.882	1.917	16.293	1.195	3.11	16.731
1.41	2.915	17.269	0.796	1.671	16.598	0.901	2.454	17.037
1.058	2.944	17.576	1.042	2.974	16.899	0.652	1.486	17.338
1.097	2.934	17.876	1.201	2.956	17.208	0.848	2.186	17.644

poor. The adjustment on the back cavity may be ascribed to a different interpretation of the epiglottis structure during the manual tracing. The largest changes have been made to [u] in the front cavity (probably due to the lack of accuracy of the plane wave model for the sublingual cavity).

Casts of the front part of the vocal tract of the subject have been made for high vowels and consonants, i.e. [i, u, f, θ, s, ʃ, ç]. These casts, made of an alginate dental impression material, were cut along the midsagittal plane in order to get the midsagittal profile, and then in the coronal planes corresponding to the grid system (that happens to be parallel lines in this part of the tract). The different slices were put on a mylar sheet and photocopied before being scanned by a video system coupled with a PC that provided midsagittal distances and cross-sectional areas. An example of the results is given in Fig. 5. The variability between the casts of the same configuration is not negligible, and the exact size of the constriction is difficult to obtain from casts, but the overall shape is consistent. However, the qualitative fit of the midsagittal distance in the vicinity of the constriction is good for [u, f, θ, s] and acceptable for [i, ʃ, ç].

6.2. Area functions

An example of area functions obtained from the corrected midsagittal functions is given in Fig. 3, and graphs of the area functions of the three cardinal vowels [a], [i] and [u] are given in Fig. 4. Since data published on area and midsagittal functions are scarce in the literature, it seemed interesting to present extensively the data obtained for the nine configurations of the corpus (see Table 3). A comparison with the area obtained in the vicinity of the constriction is shown in Fig. 5. The qualitative fit of the area function in the vicinity of the constriction is good for [i, u, f, θ, ç] and acceptable for [s, ʃ, ç]. Fig. 6(a) shows a comparison between the data on the constriction areas obtained by (1) aerodynamic measurements and (2) optimisation. Fig. 6(b) gives as well a comparison of the intralabial areas obtained by (1) measurement on front photographs, and (2) phase 2 optimisation. This figure allows to assess how the geometrical constraints have been met. It

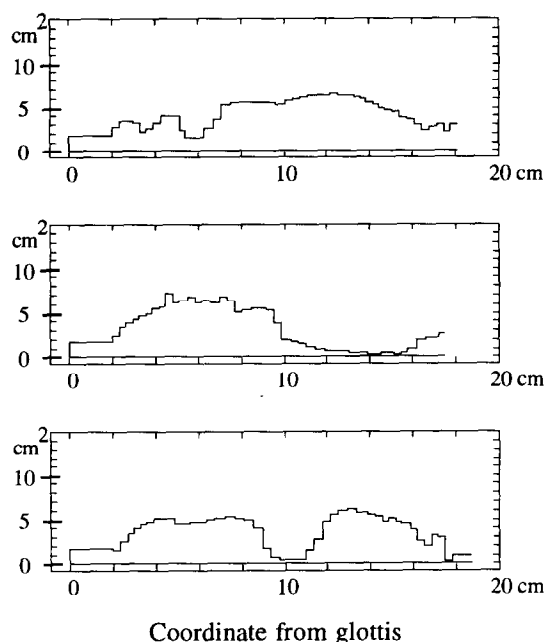


Fig. 4. Area functions for [a] (top), [i] (middle) and [u] (bottom).

can be seen that most of constriction areas are much higher than the targets. One possible explanation is that, since first formant frequencies were rather difficult to determine, they may have been attributed too high a value, leading to overly large constriction areas. It can also be seen from Fig. 6(b) that the fit for the lip areas is rather good, with exceptions for [f] (possibly due to a too low $F1$ value), and for [ʃ] for an unknown reason.

A general remark is that the vocal tract of this subject does not exhibit very large areas: the range of the maximum is about 6–7 cm², whereas some other data in the literature suggest much higher values (see (Vallée and Boë, 1992: 16 cm²; Fant, 1960: 15 cm²)). It is clear that the midsagittal distances are very rarely above 2 cm for this subject; moreover, we have filled the hard palate cast with modelling clay, sliced the result in different coronal planes, and verified that, assuming a flat tongue positioned in the plane passing by the lower edge of the upper teeth, the area was of the order of magnitude of 6–7 cm². Even for low articulations, the tongue surface does not lie below this plane in the region where the palate is the deepest.

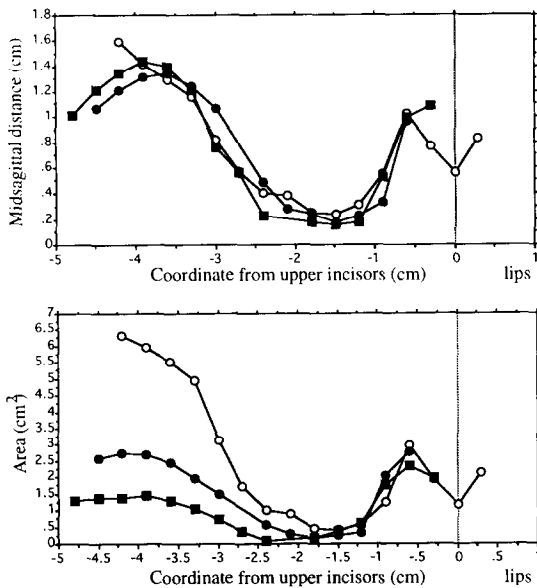


Fig. 5. Comparison of (a) midsagittal and (b) area functions, in the vicinity of the constriction, between the final configuration (open circles) and the measurements obtained from casts (filled symbols) for [s].

6.3. Coefficients α and β

The α_{inf} and α_{sup} functions are entirely defined by their Fourier series coefficients, only the three first harmonics of $\omega x = \pi x / l_{\text{tot}}$ being used. However, in order to avoid null or negative values for these functions, a minimum threshold has been set to 0.01. The Fourier coefficients are listed in Table 4, and the corresponding curves are displayed in Fig. 7a (open symbols). It is important to note that the α_{sup} curve always lies under the α_{inf} one, which means that, for a given x coordinate, $\alpha(d, x)$ decreases when d increases

Table 4

Fourier series coefficients defining the α_{inf} and α_{sup} functions versus coordinate from glottis (the first three harmonics only are used)

$\alpha_{\text{inf},0} = 2.68080$	$a_{\text{sup},0} = 0.312325$
$\alpha_{\text{inf},1} = -2.45758$	$a_{\text{sup},1} = -0.157566$
$\alpha_{\text{inf},2} = -1.45182$	$a_{\text{sup},2} = -2.26215$
$\alpha_{\text{inf},3} = 0.877827$	$a_{\text{sup},3} = -0.0180104$
$\beta_{\text{inf},1} = 1.25524$	$b_{\text{sup},1} = 0.759182$
$\beta_{\text{inf},2} = 1.90228$	$b_{\text{sup},2} = 0.137452$
$\beta_{\text{inf},3} = 0.864385$	$b_{\text{sup},3} = 1.66743$

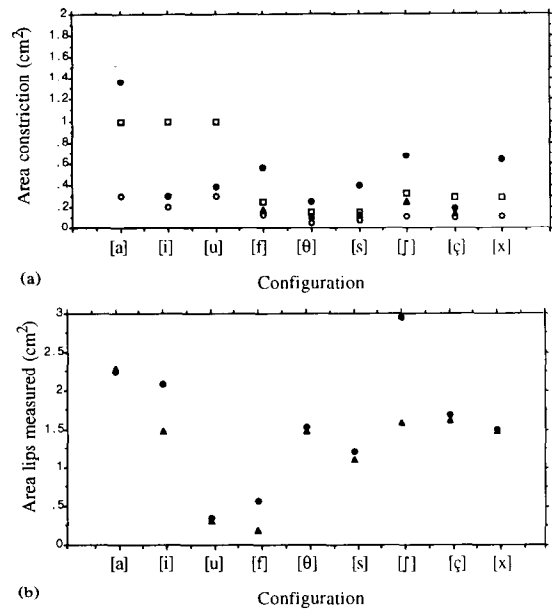


Fig. 6. Assessment of (a) constriction and (b) lip areas. (a) Area range (open symbols: $[\square]$ A_{cmin} , $[\circ]$ A_{cmax}), aerodynamic equivalent areas (filled triangles: $[\blacktriangle]$), and obtained areas (filled circles: $[\bullet]$) for the constriction. (b) The optically measured area at the lips (filled triangles: $[\blacktriangle]$) and the obtained lip areas (filled circles: $[\bullet]$).

and vice-versa, in the range between d_{inf} and d_{sup} . This would be the case if the vocal tract cross-section could be considered as a surface enclosed in a fixed length contour: a small midsagittal distance would give a small cross-sectional area, but a large midsagittal distance would imply a small lateral dimension and thus give also a small cross-sectional area. The function $A(d)$ is thus expected to attain a maximum value for a certain d_{max} . This phenomenon will be referred to as midsagittal distance/cross-sectional area non-monotonicity.

In order to assess the properties of the new model, d_{max} , $\alpha_{\text{max}} = \alpha(d_{\text{max}})$ and $A_{\text{max}} = A(\alpha_{\text{max}}, d_{\text{max}})$ have been computed for each x coordinate and plotted (see Fig. 7). To render more explicitly the behaviour of the cross-sectional area as a function of the midsagittal distance, $A(d, x)$ has been plotted versus d for three locations in the vocal tract (see Fig. 8). It can be seen (from Figs. 7 and 8) that the non-monotonicity could occur mainly in the front regions of the vocal tract,

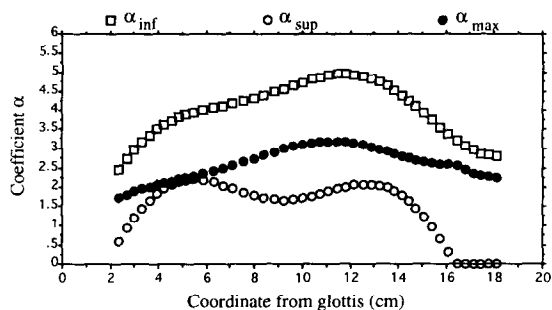


Fig. 7a. Coefficients α_{inf} (open squares), α_{sup} (open circles) and α_{max} (filled circles) versus coordinate from glottis.

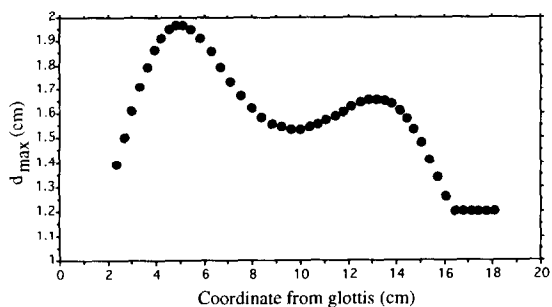


Fig. 7b. Midsagittal distance for which the cross-sectional area $A(d)$ attains a maximum, versus coordinate from glottis.

which would not be expected. In fact, it can be checked from Fig. 9, which represents the α coefficients for the nine configurations of the corpus against the section index (and not the coordinate from the glottis), that the data behave according to what one could expect: in the pharyngeal region (up to section index 22–23), the non-monotonicity phenomenon occurs in a good

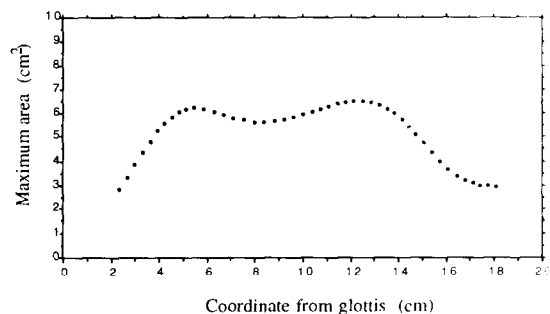


Fig. 7c. Maximum of the cross-sectional area versus coordinate from glottis.

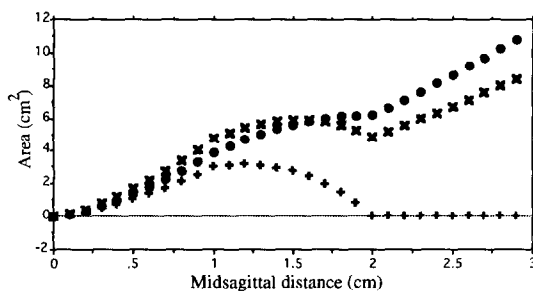


Fig. 8. Cross-sectional area versus midsagittal distance at three locations in the vocal tract ($x = 5$ cm [●], 10 cm [×], 17 cm [+]). Note that the curve marked by [+] pertains to the lip region; the unexpected values of the lip area for midsagittal distance higher than 1.5 cm are due to a bias in the model and do not correspond to physiological reality.

number of cases, whereas it never happens at the lips (in fact, over the nine configurations of the corpus, none of them has a midsagittal distance at the lips exceeding d_{max} , which means that α is always bounded by α_{inf} and α_{max}), and very little in the region around the incisors.

The unexpected aspect of the theoretical behaviour of the model for vocal tract front cavities (α_{sup} falls down to zero) can be partly explained by the fact that the corpus largely contains configurations with small midsagittal distances in this region (except for [a] and [x]). Therefore no data were available to constrain the optimisation well enough for high midsagittal distance values. As a consequence, it can be expected that this model may not be used for articulations with a rather high degree of lip aperture and for jaw opening.

It should be recalled that the β coefficient has been set to 1.5.

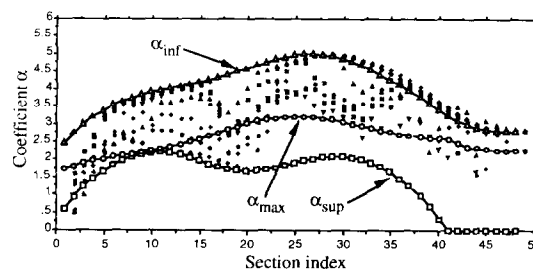


Fig. 9. Alpha coefficients versus section index for the nine configurations (the three lines have been redrawn from Fig. 7a).

Another aspect of the assessment of the α and β coefficients is a comparison with other results in the literature, and especially Sundberg and colleagues' results (Sundberg, 1969; Sundberg et al., 1987) and Baer et al. (1991).

Table 5 allows a first comparison of the α and β coefficients at the level of mean values. It appears that the value of β adopted for the complete vocal tract lies between the values for the pharynx and those for the mouth, whatever data are considered. The mean values of α are much higher in our data. There may be a compensation with β in the region of the pharynx since our β is higher than in the other studies, but this could not be the case for the region above the pharynx.

In order to compare in more detail our results with Baer et al.'s (1991) data, we have calculated the α coefficients from their initial data (see their Figs. 17–19 and 21–22), for each section, both in the pharynx and in the mouth, assuming $\beta = 1.5$ as the slopes of the regression lines of A as a function of $d^{1.5}$. These coefficients are plotted versus the distance from the teeth in Fig. 10. The comparison of Figs. 9 and 10 shows that our α coefficients are slightly larger for both pharynx and mouth regions than theirs. There are no data available in their velar region, where we get higher values of α .

6.4. Formant frequencies

The measured target formants have been given in Table 2, as well as the formants obtained after the second optimisation phase of midsagittal distance adjustment. Fig. 11 shows the relative errors obtained after both optimisation and adjustment phases. It appears that after the model optimisation phase, a number of formants have not reached correct frequencies (especially for [u, ʃ, x]). The adjustment ensures a much better fit

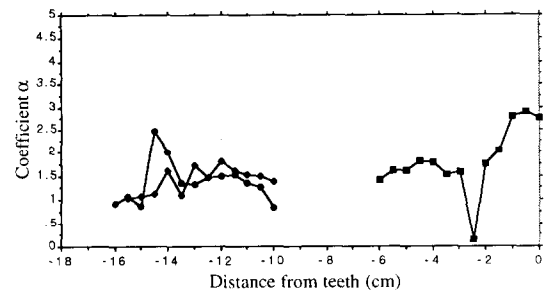


Fig. 10. Coefficients α recalculated from Baer et al.'s (1991) study by linear regression between cross-sectional areas and 1.5 powers of the midsagittal distances. The abscissa represents the distance from the teeth (these distances have been recomputed for the pharynx region), Square symbols (■) represent the pooled data for TB and PN subjects mouth, and circles (●) and diamonds (◆) represent TB and PN's pharynx data, respectively.

(3.2% error on the average), even though some errors are still too high (up to 16% for $F1$ of [i]). The high errors are related to $F1$, which could be explained partly by a higher relative measurement error (the absolute error is constant). The mean errors are 8.% for $F1$, 3.1% for $F2$ and 1.6% for $F3$.

6.5. Vocal tract length

Vocal tract lengths have already been given in Table 1. The average length for the three vowels common to Baer et al.'s data and to our data is 18 cm for subject TB (measured on X-ray scans of midsagittal profiles), 17.7 cm for PN and 18.1 for our subject PB. The length that was obtained are thus in the right range.

It should be mentioned that, in the preliminary phase of their study, an attempt had been made to adjust the length of the vocal tract by optimisation. It appeared that this was impossible, because of compensations between length and areas.

7. Conclusion and perspectives

This work aimed at two goals: (1) to derive a midsagittal profile to area function conversion model optimised for a given subject, and (2) to

Table 5
Mean values of α and β for three different studies

	Pharynx	Above pharynx
Baer et al.	$\alpha = 0.93$ $\beta = 1.75$	$\alpha = 1.27$ and $\beta = 1.40$
Sundberg	$\alpha = 1.50$ $\beta = 1.62$	$\alpha = 2.35$ and $\beta = 1.38$
This study	$\alpha = 2.98$ $\beta = 1.50$	$\alpha = 3.50$ and $\beta = 1.50$

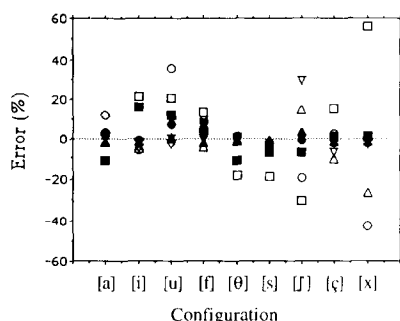


Fig. 11. Relative errors on the formants after the optimization of the model ([□] for F_1 , [○] F_2 , [△] F_3 , [▽] F_4 and [◇] F_5), and after the mid-sagittal distance adjustment ([■] for F_1 , [●] F_2 , [▲] F_3 , [▼] F_4 and [◆] F_5).

provide a set of coherent mid-sagittal functions, area functions and formant frequencies, for a small corpus of vowels and fricative consonants based on experimental data. These two goals have been reached fairly well. Indeed, a new model of mid-sagittal profile to area function has been designed as an extension of Perrier et al.'s (1992) zone model, and optimised for one subject. This model works both for vowels and fricative consonants, which seems a new point. Moreover, another set of vowels and back fricatives of Arabic for another subject has been successfully processed, which is an evidence of the generality of the proposed methodology.

However, there are still important needed improvements. A better modelling of the larynx cavity could be envisaged, as well as of the piriformis sinuses. The lip horn could be modelled in a different way by using an *acoustic equivalence*, i.e. a single cylindrical section which would have the same equivalent impedance as the lip horn; such a study is under progress. The data base should be extended to other consonant categories, such as plosives, nasals or laterals. Most of the improvement will necessitate more accurate and detailed data, that could be provided by MRI techniques, as well as ultrasound techniques associated with EPG techniques.

Even though geometric data have only partly been checked against measured data, the fact that the model has been developed with the strong constraint that it should work with the same coefficients for sounds as different as vowels and

consonants, and be coherent at both the sagittal and the acoustic level, should ensure its reliability. However, it should not be forgotten that the notion itself of area function implies implicitly an acoustic model based on plane wave propagation, and that this notion is thus limited to frequencies below 4–5 kHz. A next step in the modelling of vocal tract acoustics would be to consider true three-dimensional propagation of the acoustic waves, where it would then be unavoidable to obtain real three-dimensional vocal tract measurements.

We are currently working on the problem of the inversion of the articulatory–acoustic relationship, i.e. recovering the vocal tract shape from formants (Badin et al., Accepted). It is a well established fact that the articulatory–acoustic relationship is not univocal (see e.g. (Atal et al., 1978)), and that the ill-posed problem of acoustic inversion cannot be solved without the proper use of constraints. The new model is thus being used now as a major constraint for the inversion of non-sense words involving vowels and fricative consonants.

Acknowledgements

This work has been partly supported by the CEC in the frame of the ESPRIT/BR project SPEECH MAPS. We are deeply indebted to Dr. Pat Nye and colleagues for kindly providing the original data of the Baer et al. (1991) study. We sincerely thank Jean-Luc Schwartz, Pascal Perrier, Gérard Bailly and Bernard Gabioud for many fruitful suggestions and pertinent observations, Diane Ritterhaus for making the alginate casts and Jürgen Hess for the German translation of the abstract. We have also very much appreciated the stimulating comments of the reviewers Gunnar Fant and Juergen Schroeter.

References

- C. Abry and L.-J. Boë (1986), "Laws for lips", *Speech Communication*, Vol. 5, No. 1, pp. 97–104.
- B.S. Atal, J.J. Chang, M.V. Mathews and J.W. Tukey (1978), "Inversion of articulatory-to-acoustic transformation in the

- vocal tract by a computer-sorting technique", *J. Acoust. Soc. Amer.*, Vol. 63, pp. 1535–1555.
- P. Badin (1989), "Acoustics of voiceless fricatives: Production theory and data", *Speech Transmission Laboratory – Quarterly Progress and Status Report, KTH, Stockholm*, Vol. 3, pp. 33–55.
- P. Badin (1991), "Fricative consonants: Acoustic and X-ray measurements", *J. Phonetics*, Vol. 19, pp. 397–408.
- P. Badin and G. Fant (1984), "Notes on vocal tract computation", *Speech Transmission Laboratory – Quarterly Progress and Status Report, KTH, Stockholm*, Vols. 2/3, pp. 53–108.
- P. Badin and E. Tzavali (1990), Détermination de la fonction d'aire du conduit vocal à partir d'images téléradiographiques par rayons X [Determination of the vocal tract area function from X-ray teleradiographs], Report SC1*0147-C (EDB) to the CEC, June–December 1990, pp. 5–16.
- P. Badin, D. Beautemps, R. Laboissière and J.L. Schwartz (Accepted), "Recovery of vocal tract geometry from speech signal for vowels and fricative consonants using a midsagittal-to-area function conversion model", *J. Phonetics*.
- T. Baer, J.C. Goore, L.C. Gracco and P.W. Nye (1991), "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels", *J. Acoust. Soc. Amer.*, Vol. 90, pp. 799–828.
- A. Bothorel, P. Simon, F. Wioland and J.P. Zerling (1986), Cinéradiographie des voyelles et consonnes du français [Cineradiography of vowels and consonants in French], Travaux de l'Institut de Phonétique de Strasbourg, Strasbourg.
- J.N. Carter, T.R. Mathews and C.H. Shadle (1990), "A three-dimensional measurement system for speech research based on structured light", *Applications of Digital Image Processing*, Vol. XIII, pp. 378–387.
- E. Castelli, C. Scully and M. Castelli (1990), Avancement des travaux, Report SC1*0147-C (EDB) to the CEC, January–June 1990, pp. 35–68.
- T. Chiba and M. Kajiyama (1941), *The Vowel: Its Nature and Structure* (Tokyo-Kaiseisan Publishing, Tokyo).
- G. Fant (1960), *Acoustic Theory of Speech Production* (Mouton, The Hague).
- G. Fant (1964), "Formants and cavities", *Proc. Fifth Internat. Congress of Phonetic Sciences*, pp. 120–141.
- G. Fant (1992), "Vocal tract area functions of Swedish vowels and a new three-parameter model", *Proc. 1992 Internat. Conf. on Spoken Language Processing, Banff, Canada*, Vol. 1, Paper Fr.fAM.3.1, pp. 807–810.
- W.J. Hardcastle, F. Gibbon and K. Nicolaidis (1991), "EPG data reduction methods and their implications for studies of lingual articulation", *J. Phonetics*, Vol. 19, pp. 251–266.
- J.M. Heinz and K.N. Stevens (1965), "On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech", *Proc. Fifth Internat. Congress of Acoustics*, Paper A44.
- J.N. Holmes (1981), "Requirements for speech synthesis in the frequency range 3–4 kHz", *Proc. Fourth FASE Symposium*, pp. 169–172.
- U. Ingård (1953), "On the theory and design of acoustic resonators" *J. Acoust. Soc. Amer.*, Vol. 25, pp. 1037–1067.
- R. Laboissière (1992), Préliminaires pour une robotique de la communication parlée: inversion et contrôle d'un modèle articuloire [Preliminaries to speech robotics: inversion and control of an articulatory model], Unpublished doctoral dissertation, Institut National Polytechnique de Grenoble, France.
- M.T. Lallouache (1990), "Un poste "Visage-Parole". Acquisition et traitement de contours labiaux [A "Face-Speech" workstation. Acquisition and processing of labial contours]", *Proc. 18èmes Journées d'Etude sur la Parole, Société Française d'Acoustique*.
- M.M. Lavrentiev (1967), *Some Improperly Posed Problems of Mathematical Physics* (Springer, Berlin).
- B. Lindblom and J. Sundberg (1971), "Acoustical consequences of lip, tongue, jaw and larynx movement", *J. Acoust. Soc. Amer.*, Vol. 50, pp. 1166–1179.
- S. Maeda (1972), On the conversion of vocal tract X-ray data into formant frequencies, Bell Laboratories, Murray Hill, NJ.
- S. Maeda (1989), "Articulation compensatoire des voyelles: Analyse de données cinéradiographiques avec un modèle linéaire", in *Mélanges de Phonétique Générale et Expérimentale*, Publications de l'Institut de Phonétique de Strasbourg, pp. 545–562.
- P. Perrier, L.J. Boë and R. Sock (1992), "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients", *J. Speech Hearing Res.*, Vol. 35, pp. 53–67.
- D.E. Rumelhart, G.E. Hinton and J.L. McClelland (1986), "A general framework for parallel distributed processing", in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition 1*, ed. by D.E. Rumelhart and J.L. McClelland (MIT Press, Cambridge, MA), pp. 45–56.
- M.R. Schroeder (1967), "Determination of the geometry of the human vocal tract by acoustic measurements", *J. Acoust. Soc. Amer.*, Vol. 41, pp. 1002–1010.
- M.M. Sondhi and B. Gopinath (1971), "Determination of vocal tract shape from impulse response at the lips", *J. Acoust. Soc. Amer.*, Vol. 49, pp. 1867–1873.
- M.M. Sondhi and J.R. Resnick (1983), "The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis", *J. Acoust. Soc. Amer.*, Vol. 73, pp. 985–1002.
- M. Stone, T.H. Shawker, T.L. Talbot and A.H. Rich (1988), "Cross-sectional tongue shape during the production of vowels", *J. Acoust. Soc. Amer.*, Vol. 83, pp. 1586–1596.
- M. Stone, A. Faber, L.J. Raphael and T.H. Shawker (1991), "Cross-sectional tongue shape and linguopalatal contact patterns in [s], [ʃ], and [l]", *J. Phonetics*, Vol. 20, pp. 253–270.
- J. Sundberg (1969), "Articulatory differences between spoken and sung vowels in singers", *Speech Transmission Laboratory – Quarterly Progress and Status Report, KTH, Stockholm*, Vol. 1, pp. 33–46.

- J. Sundberg and B. Lindblom (1990), "Acoustic estimations of the front cavity in apical stops", *J. Acoust. Soc. Amer.*, Vol. 88, pp. 1313–1317.
- J. Sundberg, C. Johansson, H. Wilbrand and C. Ytterbergh (1987), "From sagittal distance to area. A study of transverse, vocal tract cross-sectional area", *Phonetica*, Vol. 44, pp. 76–90.
- J. Sundberg, B. Lindblom and J. Liljencrants (1992), "Formant frequency estimates for abruptly changing area functions: A comparison between calculations and measurements", *J. Acoust. Soc. Amer.*, Vol. 91, pp. 3478–3482.
- N. Vallée and L.J. Boë (1992), "Vers des prototypes acoustiques et articulatoires des 37 phonèmes vocaliques d'UPSID", *Proc. 19èmes Journées d'Etude sur la Parole, Société Française d'Acoustique*, pp. 53–58.
- H.Y. Wu, P. Badin, Y.M. Cheng and B. Guérin (1987), "Vocal tract simulation: Implementation of continuous variations of the length in a KELLY-LOCHBAUM model. Effects of area function spatial sampling", *IEEE Internat. Conf. Acoust. Speech Signal Process.*, Vol. 1, No. 4, pp. 9–12.