IEICE TRANS. INF. & SYST., VOL. E79–D, NO. 8 AUGUST 1996

PAPER

An Acoustically Oriented Vocal-Tract Model

Hani C. YEHIA†*, *Nonmember*, Kazuya TAKEDA†, *and* Fumitada ITAKURA†, *Members*

SUMMARY The objective of this paper is to find a parametric representation for the vocal-tract log-area function that is directly and simply related to basic acoustic characteristics of the human vocal-tract. The importance of this representation is associated with the solution of the articulatory-to-acoustic inverse problem, where a simple mapping from the articulatory space onto the acoustic space can be very useful. The method is as follows: Firstly, given a corpus of log-area functions, a parametric model is derived following a factor analysis technique. After that, the articulatory space, defined by the parametric model, is filled with approximately uniformly distributed points, and the corresponding first three formant frequencies are calculated. These formants define an acoustic space onto which the articulatory space maps. In the next step, an independent component analysis technique is used to determine acoustic and articulatory coordinate systems whose components are as independent as possible. Finally, using singular value decomposition, acoustic and articulatory coordinate systems are rotated so that each of the first three components of the articulatory space has major influence on one, and only one, component of the acoustic space. An example showing how the proposed model can be applied to the solution of the articulatory-to-acoustic inverse problem is given at the end of the paper.

key words: *vocal-tract log-area function, formant frequencies, factor analysis, independent component analysis, singular value decomposition, articulatory-to-acoustic inverse problem*

1. Introduction

In 1967, Schroeder [1] analytically described the relationship between the singularities (poles and zeros) of the vocal-tract admittance measured at the lips, and the Fourier cosine series [2] of the corresponding vocal-tract cross-sectional log-area function. The analysis was restricted to small perturbations of a uniform tract. For the case of larger variations, Mermelstein [3] developed a numerical procedure to solve the problem. It was shown that the admittance poles, which correspond to the formant frequencies, do not uniquely determine the vocal-tract log-area function. The first M formant frequencies can be used to find only the first M odd coefficients of the Fourier cosine series expansion of the log-area function. The first M even coefficients of the same Fourier series could be determined from the first M admittance zeros. However, in contrast with the formant frequencies, the admittance zeros cannot be directly extracted from the speech signal which, therefore, does not uniquely determine the corresponding log-area function.

Yehia and Itakura [4] then developed a procedure to find, among all log-area functions that can generate a given set of three formant frequencies, the one that can be reached by the vocal-tract with minimum effort. Each log-area function was parametrized by the first nine coefficients of its Fourier cosine series expansion. However, for the case of the human vocal-tract, it cannot be said that the parametrization by a truncated Fourier series is optimum. It is so because cosine functions, which are the eigenfunctions of a Fourier cosine series expansion, are "general purpose functions" that, in principle, are not directly related to the vocal-tract anatomy.

The objective of this paper is to develop a parametric representation of the vocal-tract log-area function, which is optimum from a statistical point of view and, at the same time, is suitable to study the relationship between the log-area function and the corresponding formant frequencies. Such a representation can then be used to solve the problem of finding the log-area function most likely to occur, given a set of formant frequencies.

The method used can be divided into three parts: Firstly, a factor analysis procedure is used to represent the log-area function by an appropriate number of parameters. After that, an independent component analysis technique is used to find coordinate systems, with components as independent as possible, for the articulatory space defined by the log-area parametric model; and for the acoustic space onto which it maps. Finally, singular value decomposition is used to rotate both articulatory and acoustic spaces in such a way that the first three articulatory components have major influence on one, and only one, acoustic component. These procedures are explained in detail in the next sections.

2. The Corpus

In order to find a log-area representation which is optimum from a statistical point of view, it is necessary to have a corpus of log-areas large enough to allow a good statistical characterization of the vocal-tract. The corpus used here consists of 519 log-areas. They were

Manuscript received December 7, 1995.
Manuscript revised March 4, 1996.

†The authors are with Nagoya University, Nagoya-shi, 464–01 Japan.
*Presently, with ATR Human Information Processing Res. Labs., Kyoto-fu, 619–02 Japan.

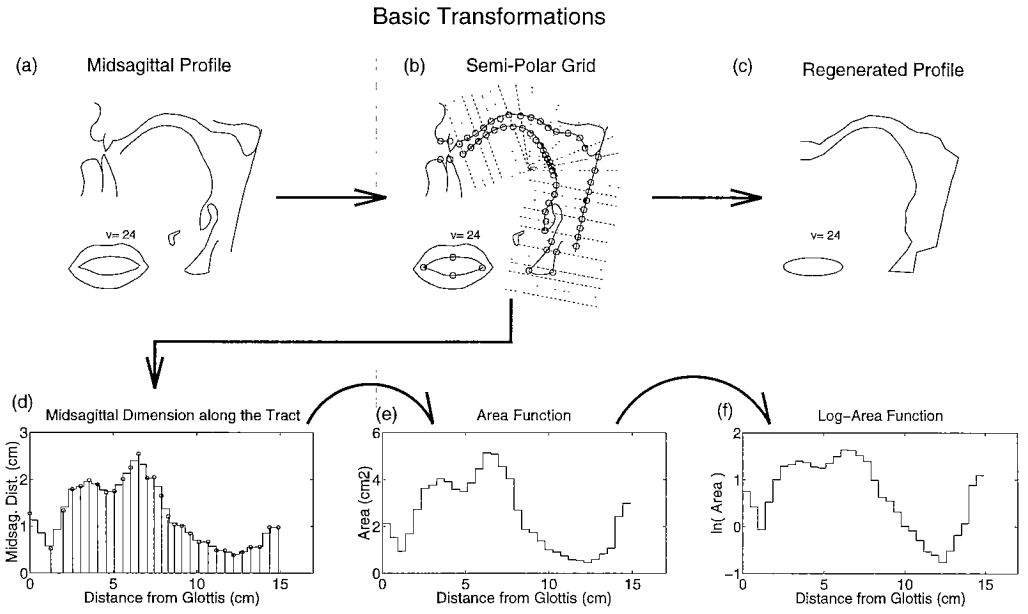


Fig. 1 Basic transformations carried out to build the corpus. (a) A sample of a midsagittal profile. (b) Semipolar grid used to sample the midsagittal dimension in 29 points. (c) Profile regenerated from the points sampled on the grid. (d) Midsagittal distances at the points sampled on the grid (stems); and corresponding evenly spaced resampled sections (stairstep graph). (e) Area function approximated by uniform tubes of same length. (f) Corresponding log-area function.

derived from midsagittal profiles obtained by cineradiography [5] from a female subject (PB). The procedure used to convert midsagittal profiles into log-areas is similar to that described by Maeda [6], and is illustrated in Fig. 1. The area functions were derived from the midsagittal dimensions using the $\alpha - \beta$ model proposed by Heinz and Stevens [7]

$$A(l) = \alpha(l) d(l)^{\beta(l)}, \tag{1}$$

where $A(l)$ and $d(l)$ are respectively the area and the midsagittal dimension at a distance l from the glottis. The values of the parameters α and β were kindly provided by Maeda who obtained them using an *ad hoc* method [6]. (Note that α and β in the equation above do not have any relation with the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that will appear later.)

Admittedly, area functions obtained from midsagittal profiles are not accurate. Even if more elaborate models, such as those proposed by Perrier et al. [8] and by Beautemps et al. [9], are used, the two-dimensional information provided by a profile is not sufficient to completely determine the area function, which depends on the three-dimensional structure of the tract. This is the main reason why the formant frequencies derived from the area function [10] do not perfectly match those extracted from the speech signal [6]. In this paper, the formant frequencies used will be always derived from the area function, using the method proposed in [10]; and not the formants extracted from the speech signal. By doing so, any problems due to inaccuracies inherent in the area function estimation method are avoided.

The log-area function is simply the natural logarithm of the area function. The only detail is that, in order to avoid numerical problems with closures, areas smaller than a given threshold ϵ (in this paper, $\epsilon = 5\,\text{mm}^2$) are clipped to ϵ. It, however, does not lead to any considerable inaccuracies from either articulatory or acoustic points of view.

Each log-area function present in the corpus, when approximated by a concatenation of uniform tubes of equal length, as in the example shown in Fig. 1 (f), can be represented by a vector containing the natural logarithm of the section areas and the tract length. In this paper, the following notation will be used

$$\begin{aligned} \mathbf{x}_i &= [x_{1i}, \ldots, x_{Ki}, x_{K+1i}]^t, \quad i = 1, \ldots, P, \\ x_{ki} &= \ln A_i \left[\left(k - \frac{1}{2} \right) \frac{L_i}{K} \right], \quad k = 1, \ldots, K, \\ x_{K+1i} &= L_i, \end{aligned} \tag{2}$$

where L_i is the tract length of frame i, expressed in units normalized so that the variance of x_{K+1} is equal to the largest variance of the first K components of $\mathbf{x}$; $A_i(l)$ is the cross-section area of frame i at distance l from the glottis; $K = 32$ is the number of uniform sections present in each area function; and $P = 519$ is the number of vectors present in the corpus.

3. The Parametric Model

The objective of this section is to find representations for both the log-area space and the formant frequency space so that

- Each space be efficiently represented by a small number of parameters.
- The components of each space be as independent as possible.
- The mapping between both spaces be as simple as possible.

These points will be analyzed one by one in the following sections.

3.1 Eigenvalue Decomposition

3.1.1 Articulatory Space

The number of sections necessary to obtain a good approximation of the vocal-tract log-area function by a concatenation of uniform tubes of equal length is considerably larger than the dimension of the space composed by the log-area functions that can be produced by the human vocal-tract. This space, from now on, will be called the *articulatory space,* and an eigenvalue decomposition procedure will be carried out to parametrize it by an appropriate number of components.

The procedure is as follows: Given the corpus of log-area vectors defined in Eq. (2), the corresponding covariance matrix is given by

$$\mathbf{C} = \frac{1}{P-1}\sum_{i=1}^{P}[\mathbf{x}_i - \boldsymbol{\mu}_x][\mathbf{x}_i - \boldsymbol{\mu}_x]^t, \tag{3}$$

where $\boldsymbol{\mu}_x$ is the mean log-area vector; and can be expressed as

$$\mathbf{C} = \mathbf{U}\mathbf{S}\mathbf{U}^t, \tag{4}$$

where $\mathbf{S}$ is a diagonal matrix containing the eigenvalues of $\mathbf{C}$ in decreasing order, and $\mathbf{U}$ is a unitary matrix whose columns contain the corresponding normalized eigenvectors. The expansion above is a *Takagi's factorization,* which is a *singular value decomposition* for the particular case of symmetric matrices [11].

Using the same optimality principle of the Karhunen-Loève transform [12], $\mathbf{x}$ can then be approximated by

$$\mathbf{x} \approx \mathbf{U}_N\boldsymbol{\alpha} + \boldsymbol{\mu}_x, \tag{5}$$

$\boldsymbol{\alpha}$ given by

$$\boldsymbol{\alpha} = \mathbf{U}_N^t(\mathbf{x} - \boldsymbol{\mu}_x), \tag{6}$$

where $\mathbf{U}_N$ is the matrix containing the first N columns of $\mathbf{U}$, i.e. the normalized eigenvectors corresponding to the N largest eigenvalues of $\mathbf{C}$. The $K+1 = 33$ eigenvalues are shown in Fig. 2. Note that only the first $N = 5$ eigenvalues have non-negligible values, and that they "explain" more than 92% of the variance of the corpus of log-areas.

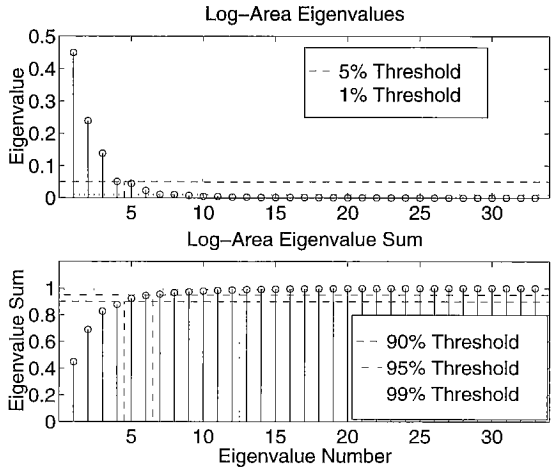


Fig. 2 Eigenvalues of the log-area covariance matrix.

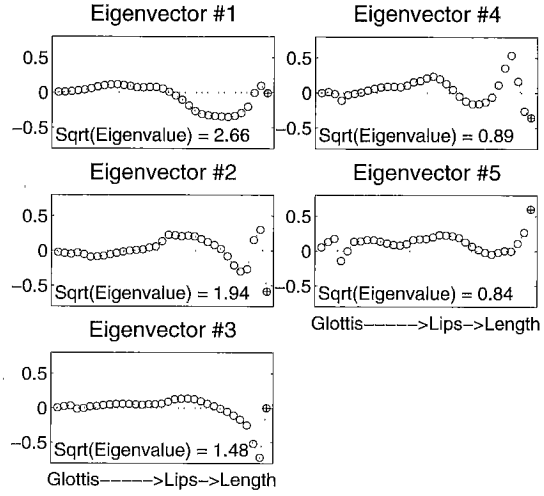


Fig. 3 Eigenvectors corresponding to the first 5 eigenvalues obtained from the decomposition of the log-area covariance matrix. All eigenvectors are normalized to have unit Euclidian norm. The first $K = 32$ components correspond to the log-area along the tract; and the last component corresponds to the tract length. The corresponding eigenvalue square root is given as a reference to the "importance" of each eigenvector.

The eigenvectors associated with the largest $N = 5$ eigenvalues are shown in Fig. 3. They will be used in this paper to form a *parametric model* for the vocal-tract log-area function. Since the components of this model cannot be explicitly interpreted as articulators, it cannot be qualified as an articulatory model [6], [13]. In spite of that, it is possible to observe in Fig. 3 that: the first and most important eigenvector is associated with the tongue region; the tract-length is the dominant component of the second and fifth eigenvectors; the lips determine the dominant component of the third eigenvector; and the tongue apex is the dominant region of the fourth eigenvector. Also, note that there is almost no influence of the glottal region on the first three eigenvectors.

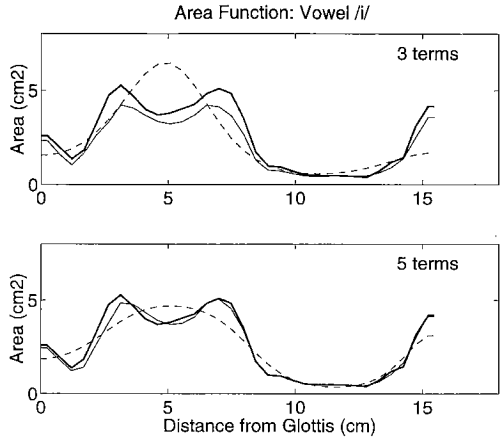


Fig. 4 Area function approximations by Fourier cosine series expansion (dashed line), and by statistically optimum eigenvalue expansion (solid line). The thick solid line shows the original area. Above: expansion with 3 components. Below: expansion with 5 components.

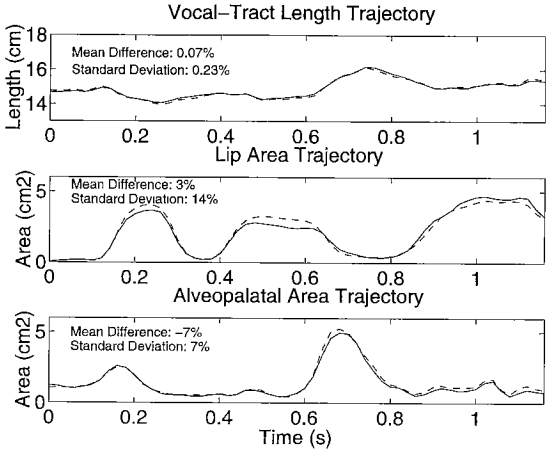


Fig. 5 Vocal-tract length, lip area, and alveopalatal area trajectories along the sentence (in French): *Ma chemise est roussie.* The dashed lines show the original measured trajectories, while the solid lines show the trajectories parametrized by the model proposed here. For each case, the mean and the standard deviation values of the relative difference (in *percentage*) between parametrized and original trajectories are also shown.

In order to illustrate the performance of this representation, Fig. 4 shows an area function taken from the corpus (thick line), and its approximations by a truncated Fourier cosine series [4] (dashed line) and by the parametric model proposed here (thin solid line). Note that, in contrast with the Fourier series representation, the parametric model is able to "capture" the vocal-tract structure. In Fig. 5, original trajectories followed by the tract length, by the area at the lips, and by the area of a section in the alveopalatal region are shown by the dashed lines. The corresponding trajectories obtained with the parametric model proposed here are shown by the solid lines. Since the parametric model is derived from the log-area function, the approximation is particularly good for small areas, which are critical from the acoustic point of view.

Summarizing, it was shown that vocal-tract log-area vectors can be efficiently represented in an $N = 5$ dimensional articulatory space. Here, it is interesting to note that most articulatory models [6], [13] are expressed by seven to nine components. This happens because their formulation is oriented to the speech production direct problem. In that case, it is important to consider the number of *degrees of freedom* of the vocal apparatus, which is usually larger than the *dimension* of the articulatory space.

3.1.2 Acoustic Space

To each log-area vector there exist one, and only one, set of formant frequencies associated with it. Here, the set composed by the first three formant frequencies will be called a *formant vector*, and the space formed by all formant vectors that can be generated by the vocal-tract will be called the *acoustic space*.

By performing an eigenvalue decomposition on the covariance matrix of the formant vectors (in log-scale), it was found that more than 92% of the total variance can be explained by the first two eigenvalues. For this reason, the possibility of representing the acoustic space in two dimensions was considered. However, since the acoustic information associated with the third eigenvalue can be important for the inverse problem, it was decided to use the first three formant frequencies to parametrize a three-dimensional acoustic space.

3.2 Independent Component Analysis

The objective of this section is to perform linear transformations on the coordinate systems of both articulatory and acoustic spaces, so that the components of each space become as independent as possible. The final objective is to find a mapping of the articulatory space onto the acoustic space, where each component of the acoustic space is mainly determined by one, and only one, component of the articulatory space. Also, each component of the articulatory space must have major influence on at most one component of the acoustic space. In order to attain this objective, a necessary condition is that the components of each space be as independent as possible.

3.2.1 Articulatory Space

The first step is to find how the articulatory space, defined in the last section, maps onto the acoustic space. To reach this target, firstly, the hyperrectangle defined by the maximum and minimum points of each of the $N = 5$ components of the parametrized corpus is "filled" with $Q_0 = 30,000$ uniformly distributed points. Figure 6 (a) illustrates this operation by showing the projection on the subspace defined by α_1 and α_2. However, not all

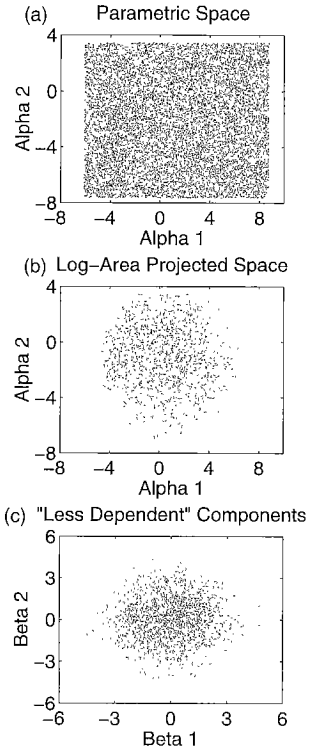


Fig. 6 (a) Parametric subspace determined by the first two components of α. (b) Points corresponding to realistic area functions (articulatory space). (c) The same points shown in a coordinate system with "less dependent" components.

the points in the hyperrectangle correspond to realistic vocal-tract areas. For this reason, all points in the hyperrectangle that correspond to areas out of the limits defined by the $P = 519$ areas present in the corpus are discarded. The remaining $Q = 7,285$ points are shown in Fig. 6 (b). After that, the *independent component analysis* method proposed by Bell [14] is applied to these points to find a linear transformation ($\mathbf{T}_{\alpha\beta} : \mathbb{R}^5 \rightarrow \mathbb{R}^5$) that changes the coordinate system of the articulatory space into a system with statistically "less dependent" components. (The term "less dependent" is used because, in the present case, a simple linear transformation is not enough to obtain a complete decomposition into independent components.) Mathematically, this transformation is written as

$$\boldsymbol{\beta} = \mathbf{T}_{\alpha\beta}\,(\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha), \tag{7}$$

where $\boldsymbol{\mu}_\alpha$ is the mean of the $Q = 7,285$ vectors generated to "fill" the articulatory space. Figure 6 (c) shows the same points shown in Fig. 6 (b), now plotted in the new coordinate system.

3.2.2 Acoustic Space

For a given point $\boldsymbol{\beta}$ in the articulatory space, it is possible to find the corresponding log-area vector $\mathbf{x}$ using the following inverse transformation

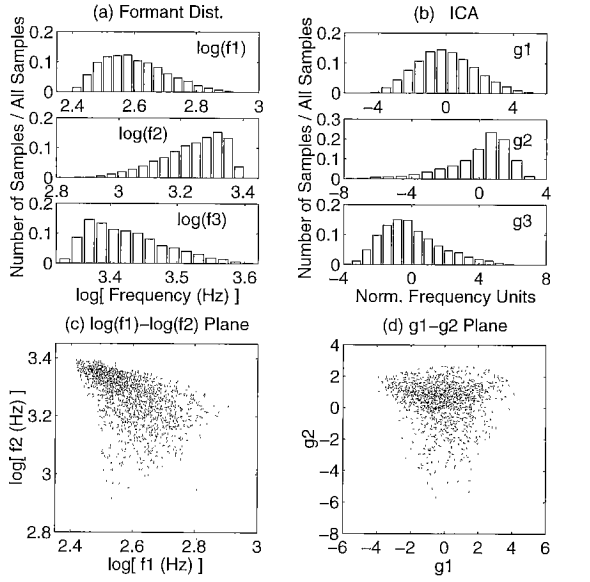


Fig. 7 (a) Normalized histograms of the first 3 formant log-frequencies corresponding to an articulatory space filled with approximately uniformly distributed points. (b) Histograms of the variables obtained after independent component analysis (ICA) of the formant frequencies. (c) and (d) Scatter plots of the first 2 variables shown in (a) and (b), respectively.

$$\mathbf{x} = \mathbf{U}_N(\mathbf{T}_{\alpha\beta}^{-1}\boldsymbol{\beta} + \boldsymbol{\mu}_\alpha) + \boldsymbol{\mu}_x. \tag{8}$$

Then, using the wave propagation model described in [10], it is possible to calculate the formant vector $\mathbf{f}$ formed by the first 3 formant frequencies associated with $\mathbf{x}$ and, consequently, with $\boldsymbol{\beta}$

$$\mathbf{f} = \mathbf{f}(\boldsymbol{\beta}). \tag{9}$$

This procedure was carried out for all $Q = 7,285$ points shown in Fig. 6 (c). The corresponding formant log-frequency normalized histograms, which are approximations for the probability density functions, are shown in Fig. 7 (a); while the scattering on the plane defined by $\log(f_1)$ and $\log(f_2)$ is shown in Fig. 7 (c).

After that, the independent component analysis (ICA) method described in [14] was used to find a linear transformation ($\mathbf{T}_{fg} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$) that changes the coordinate system defined by the formant log-frequencies into a system with "less dependent" variables. This transformation can be written as

$$\mathbf{g} = \mathbf{T}_{fg}\,[\log(\mathbf{f}) - \boldsymbol{\mu}_{\log f}], \tag{10}$$

where $\boldsymbol{\mu}_{\log f}$ is the mean of the logarithm of the $Q = 7,285$ formant vectors available. The normalized histograms obtained for the components of $\mathbf{g}$ are shown in Fig. 7 (b), and the scattering of the first two components of $\mathbf{g}$ is shown in Fig. 7 (d).

At this point, $\mathbf{g}$ and $\boldsymbol{\beta}$ define respectively acoustic and articulatory vector variables whose components are more independent than the components of $\mathbf{f}$ and $\boldsymbol{\alpha}$. The next step is to model the relationship between acoustic and articulatory spaces.

Before continuing, it is worthwhile to write some lines about the *independent component analysis* (ICA) technique used here. The ICA problem consists of finding a linear transformation which, when applied to a given ensemble of random vectors, transforms it into an ensemble of vectors whose components are statistically independent, in an ideal case; or as independent as possible, in practical cases. The approach described in [14] (and used in this paper) is based on entropy maximization which, under appropriate conditions, implies mutual information minimization, and consequent independence maximization. The method was originally used to solve the problem of blind separation of mixed sound sources, but has a potentially larger range of applications.

3.3 Singular Value Decomposition

In this section, the mapping from $\boldsymbol{\beta}$ onto $\mathbf{g}$ is approximated by a linear transformation ($\mathbf{M} : \mathbb{R}^5 \rightarrow \mathbb{R}^3$) as follows

$$\mathbf{g} \approx \mathbf{M}\boldsymbol{\beta}. \tag{11}$$

In such a case, once there is an ensemble of vectors $\mathbf{g}$ and $\boldsymbol{\beta}$ available, a minimum mean square error (MMSE) procedure can be used to estimate $\mathbf{M}$, yielding

$$\mathbf{M} = \mathbf{G}\mathbf{B}^t(\mathbf{B}\mathbf{B}^t)^{-1}, \tag{12}$$

with

$$\mathbf{G} = [\mathbf{g}_1 \dots \mathbf{g}_Q], \tag{13}$$

and

$$\mathbf{B} = [\boldsymbol{\beta}_1 \dots \boldsymbol{\beta}_Q]. \tag{14}$$

In the above equations, $Q = 7,285$ is the number of points present in the ensembles.

Once $\mathbf{M}$ is determined, a singular value decomposition procedure [11] can be used to find rotations of the acoustic ($\mathbf{g}$) and articulatory ($\boldsymbol{\beta}$) coordinate systems, so that each of the first three components of the articulatory space has major influence on one, and only one, component of the acoustic space. The singular value decomposition of $\mathbf{M}$ yields

$$\mathbf{M} = \mathbf{U}_{gh}\boldsymbol{\mu}\mathbf{U}_{\beta\gamma}^t, \tag{15}$$

where $\mathbf{U}_{gh}$ is a unitary matrix containing the normalized eigenvectors of $\mathbf{M}\mathbf{M}^t$, $\mathbf{U}_{\beta\gamma}$ is a unitary matrix containing the normalized eigenvectors of $\mathbf{M}^t\mathbf{M}$, and $\boldsymbol{\mu}$ is a 3×5 matrix whose first 3 columns define a diagonal matrix containing the square roots of the eigenvalues of $\mathbf{M}\mathbf{M}^t$, and the elements of the last two columns are all equal to zero. Now, since the multiplication of an unitary matrix by a vector represents a rotation of this vector,

$$\boldsymbol{\gamma} = \mathbf{U}_{\beta\gamma}^t\boldsymbol{\beta} \tag{16}$$

and

$$\mathbf{h} = \mathbf{U}_{gh}^t\mathbf{g} \tag{17}$$

define, respectively, "rotated" articulatory and acoustic variables. The corresponding matrix of correlation coefficients [17] can be estimated by

$$\mathbf{R} = \frac{\mathbf{H}\Gamma^t}{(Q-1)\boldsymbol{\sigma}_h\boldsymbol{\sigma}_\gamma{}^t}, \tag{18}$$

where

$$\mathbf{H} = [\mathbf{h}_1 \dots \mathbf{h}_Q], \tag{19}$$

$$\Gamma = [\boldsymbol{\gamma}_1 \dots \boldsymbol{\gamma}_Q], \tag{20}$$

$\boldsymbol{\sigma}_h$ and $\boldsymbol{\sigma}_\gamma$ are the column vectors containing respectively the standard deviations of $\mathbf{h}$ and $\boldsymbol{\gamma}$, $Q = 7,285$ is the number of points present in the ensembles, and the division of $\mathbf{H}\Gamma^t$ by $\boldsymbol{\sigma}_h\boldsymbol{\sigma}_\gamma{}^t$ is performed element-wise. The numerical result obtained is shown below

$$\mathbf{R} = \begin{bmatrix} 0.939 & 0.003 & 0.005 & -0.004 & -0.004 \\ 0.003 & 0.953 & 0.003 & -0.004 & 0.002 \\ 0.002 & 0.001 & 0.461 & 0.001 & 0.002 \end{bmatrix}.$$

This matrix shows that there exists a high degree of correlation between the first two acoustic components and the first two articulatory components. There is also a not negligible degree of correlation between the third acoustic and articulatory components. All other correlation coefficients are very small.

At this point, in order to see the importance of the independent component analysis described in Sect. 3.2, it is interesting to compare $\mathbf{R}$ with the matrix of correlation coefficients obtained when $\mathbf{f}$ and $\boldsymbol{\alpha}$ are used in place of $\mathbf{g}$ and $\boldsymbol{\beta}$ to obtain $\mathbf{h}$ and $\boldsymbol{\gamma}$, as done in [15], [16]. The result is shown below

$$\mathbf{R}_{f\alpha} = \begin{bmatrix} 0.944 & -0.270 & -0.206 & -0.308 & -0.035 \\ -0.270 & 0.944 & -0.266 & 0.258 & 0.183 \\ -0.112 & -0.142 & 0.511 & -0.069 & -0.184 \end{bmatrix}.$$

Note that, although the correlation between the acoustic components and the corresponding first three articulatory components continues to exist, the other correlation coefficients are not negligible any more.

It should be pointed out, however, that uncorrelation does not imply independence. This fact is illustrated in Fig. 8, where scatterings representing the joint cross-distributions of the components of $\mathbf{h}$ and of $\boldsymbol{\gamma}$ are plotted. There exists, for example, an apparent nonlinear relation between h_3 and γ_1. This kind of dependence cannot be well approximated by the linear transformation used in this work to model the mapping from the articulatory space onto the acoustic space.

In spite of these limitations, the model successfully extracted two acoustic variables, namely h_1 and h_2, which depend approximately linearly on two, and only two, articulatory variables, namely γ_1 and γ_2. The

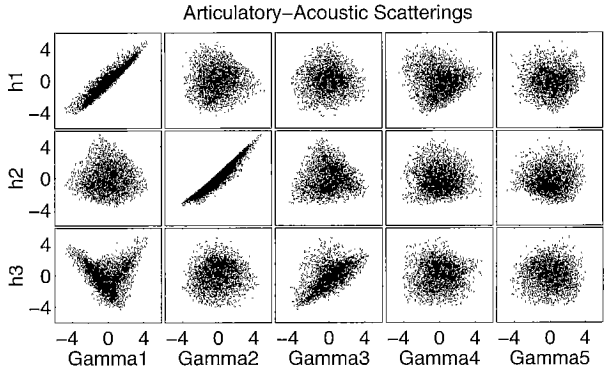


Fig. 8 Scatterings representing the joint distributions of the components of the acoustic variable $\mathbf{h}$ and the components of the articulatory variable γ. Note the high correlation between γ_1 and h_1, and between γ_2 and h_2. See also the nonlinear relation between γ_1 and h_3.

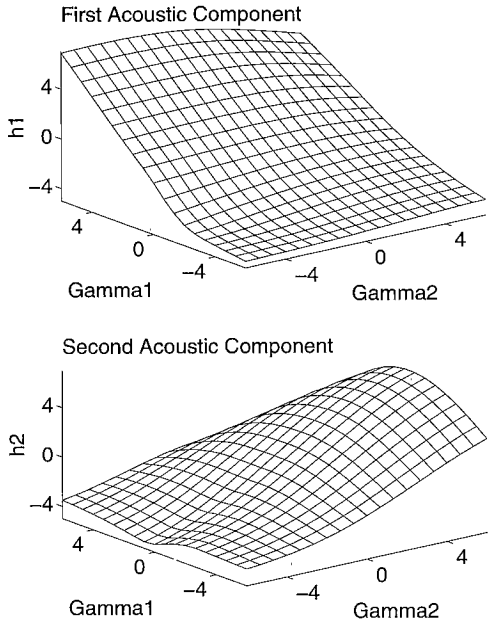


Fig. 9 First two acoustic components (h_1, and h_2) expressed as functions of the first two articulatory components (γ_1 and γ_2), when all other components (γ_3, γ_4, and γ_5) are equal to zero. Note that h_1 is almost independent of γ_2, and that there are one-to-one relationships between h_1 and γ_1, and between h_2 and γ_2.

remaining articulatory components, γ_3, γ_4, and γ_5, have little influence on h_1 and h_2. Moreover, γ_2 has little effect on h_1, and the influence of γ_1 on h_2 does not affect the one-to-one relationship between γ_2 and h_2. These facts are illustrated in Fig. 9.

Once the parametric model is derived, and its basic characteristics are analyzed, it is interesting to compare articulatory and acoustic component trajectories for a given sequence of vocal-tract shapes. The trajectories associated with the French sentence *"Ma chemise est roussie"* are shown in Fig. 10. It is possible to observe that the first two articulatory components are indeed

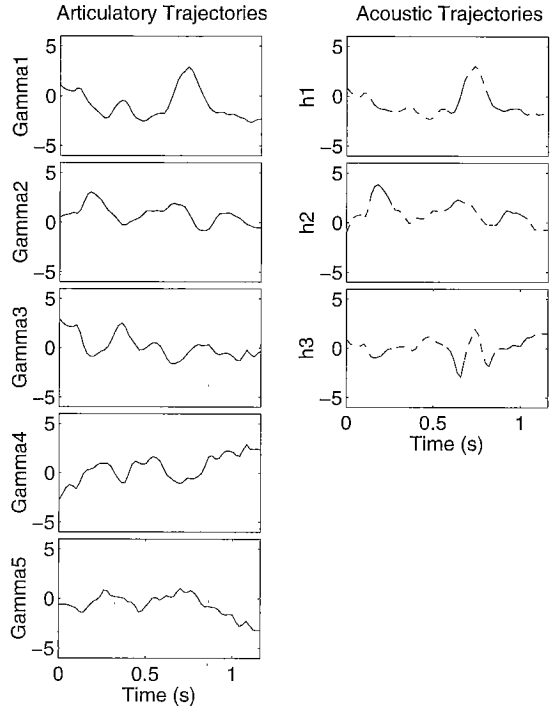


Fig. 10 Articulatory and acoustic component trajectories along the sentence (in French): *Ma chemise est roussie.* Note the similarity between the first two articulatory trajectories and the first two acoustic trajectories. (The dashed lines in the acoustic trajectories indicate the intervals where the formants cannot be reliably extracted from the speech signal due to very narrow constrictions in the area function.)

closely related to the first two acoustic components. It is also possible to see that there exist some similarities between h_3 and h_1, indicating that they are not independent.

4. Application: The Inverse Problem

The main motivation that led to the construction of the model described here is the solution of the *articulatory-to-acoustic inverse problem.* The idea used is that, if it is possible to find a simple relation between acoustic and articulatory parameters, then it is possible to represent acoustic constraints in the articulatory space, and combine them directly with minimum effort and continuity constraints. Such a combination is necessary once acoustic constraints do not uniquely determine the vocal-tract geometry [3], [4], [18].

In [19] and [4] the vocal-tract log-area function was parametrized by a truncated Fourier cosine series. After that, the acoustic constraint imposed by the first three formant frequencies was combined with minimum effort constraints expressed by a quadratic cost function. Here, instead of a Fourier series, the parametric model described in Sect. 3 was used to represent the vocal-tract. Then, using the method to combine acoustic and anatomical information described in [4], and

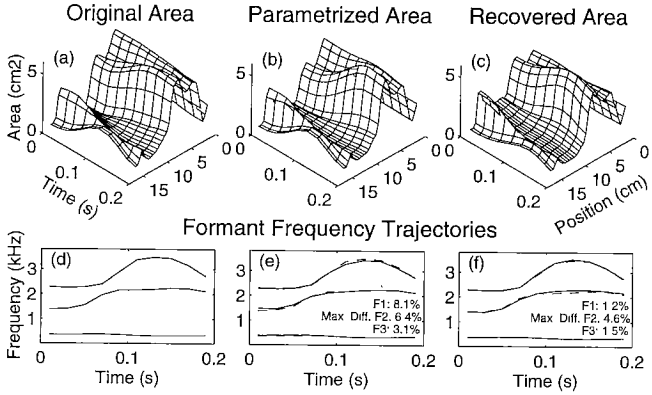


Fig. 11 (a) Sequence of area functions, taken from the corpus, corresponding to the diphthong /ui/, uttered in the (French) sentence "Luis pense à ça." (b) Sequence of areas reconstructed from the parametric representation of the original areas shown in (a). (c) Sequence of areas estimated from the formant trajectories shown in (d), under continuity and minimum effort constraints. (d), (e) and (f) Formant frequency trajectories corresponding to the sequences of areas shown in (a), (b) and (c) respectively. The dashed lines shown in (e) and (f) are the original formant trajectories shown in (d). For each pair of formant trajectories, the maximum relative difference (in *percentage*) is also shown.

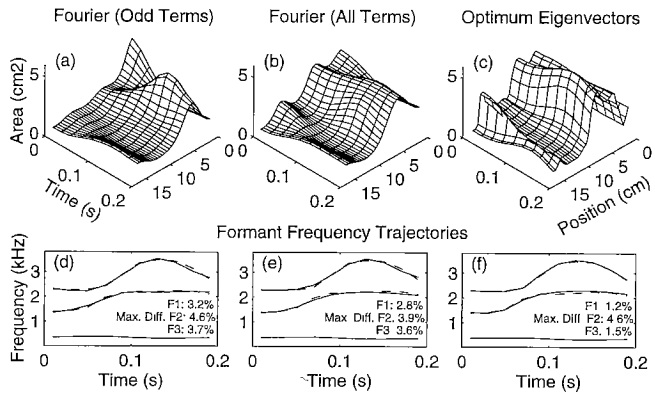


Fig. 12 Above: Sequences of area functions estimated from the formant trajectories shown in Fig. 11 (d) under different constraints: (a) The areas are represented by the first 6 components of its Fourier cosine series expansion. The even coefficients are set to zero. (b) Same as (a), but the 6 coefficients are determined under anatomical constraints. (c) The same anatomical constraints of (b) are used, but with optimized eigenvectors in place of the cosine functions used in the Fourier series expansion. Below: The solid lines in (d), (e) and (f) show the formant trajectories associated with the sequences of areas shown in (a), (b) and (c), respectively. The dashed lines are the original formant trajectories shown in Fig. 11 (d). For each pair of formant trajectories, the maximum relative difference (in *percentage*) is also shown.

the continuity constraints explained in [19], it was possible to estimate sequences of area functions from the corresponding first three formant trajectories. (A detailed analysis of this method will be left for a future publication, because it is not the target of this paper.)

In the example given in Fig. 11, the sequence of area functions shown in (a) was used to generate the formant trajectories shown in (d). These trajectories were then used to recover the original sequence of areas, under minimum effort and continuity constraints. The result is shown in (c). The search for the best sequence of areas was performed in the articulatory "γ" space. Note, however, that the sequence of areas shown in (c) is close, but not identical, to that shown in (b), which is the sequence of areas reconstructed from the parametric "γ" representation of the sequence shown in (a). The reason for this is probably associated with the fact that the mathematical cost function used does not perfectly reflect the articulation effort determined by the human physiology.

For comparison purposes, it is interesting to see the results when the same problem is solved using a truncated Fourier series to represent the log-area function, as done in [3], [19] and [4], instead of the model described in this paper. In [3], Mermelstein parametrized the vocal-tract log-area function by the first six coefficients of its Fourier cosine series expansion. It was verified that, when the even coefficients are all equal to zero, there exist a one-to-one relationship between the first three formants and the three odd Fourier coefficients. Using this property, an interactive procedure was implemented to find the unique set of odd Fourier coefficients associated with a given set of formant frequencies, when all even Fourier coefficients are equal to zero. This procedure was used to obtain the sequence of areas shown in Fig. 12 (a) from the formant trajectories shown in Fig. 11 (d). Note that the result is substantially different from the original sequence of areas shown in Fig. 11 (a). This happens because setting all even Fourier coefficients to zero is an artificial constraint that does not reflect the geometrical constraints determined by the vocal-tract anatomy. A mathematical framework to incorporate such *anatomical constraints* into the articulatory-to-acoustic inverse problem was proposed in [19] and [4]. This method was used to obtain the sequence of areas shown in Fig.12 (b). It can be seen that it resembles the original sequence of areas shown in Fig. 11 (a). However, abrupt variations, inherent in some regions of the vocal-tract, cannot be well approximated, due to the smooth character of the cosine functions, which are the eigenvectors of the Fourier cosine series representation. This is in contrast with the eigenvectors used in the present paper (see Fig. 3, which allow a good representation of the vocal-tract structure. When such eigenvectors are used in place of cosine functions, the result obtained is the sequence of areas shown in Fig.11 (c), which is reproduced in Fig.12 (c) to allow a better comparison.

The formant frequency trajectories associated with the sequences of areas shown in Fig.12 (a), (b) and (c) are shown respectively by the solid lines in Fig. 12 (d), (e) and (f). The dashed lines represent the original formant trajectories shown in Fig. 11 (d). These figures illustrate the fact that, even under continuity constraints,

substantially different sequences of area functions can generate basically the same formant trajectories.

5. Conclusion

In this paper, an acoustically oriented vocal-tract parametric model was described. In contrast with articulatory models, which have the objective of representing the vocal-tract in terms of elementary *articulators*, the elements of the model presented here are approximately linearly related to basic *acoustic* characteristics of the vocal-tract.

The vocal-tract geometry was represented by vectors containing the vocal-tract length and the natural logarithm of the cross-sectional area, sampled at 32 points evenly spaced along the tract. A factor analysis technique was then used to find an appropriate number of dimensions, namely five, for the *articulatory space*. The acoustic characteristics of the vocal-tract were represented by vectors containing the logarithms of the first three formant frequencies. The set of all formant vectors that can be generated by all possible articulatory vectors define the *acoustic space*. The articulatory space maps onto the acoustic space.

It was verified that, when appropriate linear transformations are applied to the coordinate systems of both articulatory and acoustic spaces, the first two articulatory components are highly correlated with the first two acoustic components. There is also a relatively high correlation between the third articulatory and acoustic components. The influence of the fourth and fifth articulatory components on the acoustic components is small.

An important application of the parametric model described here is in the solution of the *articulatory-to-acoustic inverse problem*. The fact that there exists an almost linear relationship between acoustic and articulatory variables allows a simple and efficient formulation for the combination of acoustic, minimum effort, and continuity constraints. An example was given for the case of a diphthong, indicating that the method can be successfully used.

Although the results obtained can be considered satisfactory, some problems still remain to be solved. Firstly, it is important to form a corpus with more accurately measured areas. Secondly, the nonlinear relations observed must be better analyzed, since they cannot be modelled only by linear transformations. Also important is the problem of speaker adaptation: The vocal-tract parametric model presented here was derived for a particular speaker using a data dependent technique. So, when the model is used, for example, in the solution of the inverse problem with a different speaker, an adaptation procedure is required. Finally, if the corpus of areas available is large enough, it would be interesting to expand the idea of a vocal-tract *position* parametric model, to a vocal-tract *gesture* parametric model. It could be important to allow a better modelling of continuity constraints during the speech production process.

References

[1] M.R. Schroeder, "Determination of the geometry of the human vocal-tract by acoustical measurements," Journal of the Acoustical Society of America, vol.41, no.4, pp.1002–1010, 1967.
[2] H.F. Davis, "Fourier Series and Orthogonal Functions," pp.107–112, Dover, 1963.
[3] P. Mermelstein, "Determination of vocal-tract shape from measured formant frequencies," Journal of the Acoustical Society of America, vol.41, no.5, pp.1283–1294, 1967.
[4] H. Yehia and F. Itakura, "A method to combine acoustical and morphological constraints in the speech production inverse problem," to appear in Speech Communication, vol.18, no.2, 1996.
[5] A. Bothorel, P. Simon, F. Wioland, and J. P. Zerling, "Cinéradiographie de voyelles et consonnes du Français," Institut de Phonétique de Strasburg, 1986.
[6] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in Speech Production and Speech Modelling, eds. W.J. Hardcastle and A. Marchal, pp.131–149, Kluwer Academic Publishers, 1990.
[7] J.M. Heinz and K.N. Stevens, "On the derivation of area functions and acoustic spectra from cineradiographic films of speech," Journal of the Acoustical Society of America, vol.36, p. 1037, 1964.
[8] P. Perrier, L.J. Boe, and R. Sock, "Vocal tract area function estimation from midsagittal dimensions with ct scans and a vocal tract cast: modelling the transition with two sets of coefficients," Journal of Speech and Hearing Research, vol.35, pp.53–67, 1992.
[9] D. Beautemps, P. Badin, and R. Laboissière, "Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data," Speech Communication, vol.16, pp.27–47, 1995.
[10] M.M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.35, no.7, pp.955–967, 1987.
[11] R. Horn and C. Johnson, "Matrix Analysis," pp.414–420, Cambridge, 1985.
[12] N. Jayant, "Digital Coding of Waveforms," pp.535–546, Springer-Verlag, 1984.
[13] C. Coker, "A model of articulatory dynamics and control," Proc. IEEE, vol.64, no.4, pp.452–460, 1976.
[14] A. Bell and J. Seijnowski, "Blind separation and blind deconvolution: An information-theoretic approach," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 1995.
[15] H. Yehia and F. Itakura, "Combining dynamic and acoustic constraints in the speech production inverse problem," IEICE Technical Report, SP95-13, 1995.
[16] H. Yehia, K. Takeda, and F. Itakura, "A vocal-tract area function trajectory representation oriented to the speech production inverse problem," Proc. 1995 autumn meeting of the Acoustical Society of Japan, pp.339–340, 1995.
[17] A. Papoulis, "Probability, Random Variables, and Stochastic Processes," McGraw-Hill, p.152, 1991.
[18] B.S. Atal, J.J. Chang, and J.W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal-tract by a computing sorting technique," Journal of the Acous-

tical Society of America, vol.63, no.5, pp.1535–1555, 1978.

[19] H. Yehia and F. Itakura, "Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal fourier analysis," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.477–480, 1994.

Appendix: Numeric Information

Most of the numeric information used in the implementation of the vocal-tract parametric model described in this paper was not included in the main text. Instead, for practical purposes, it is given in this appendix, and can be used by the interested reader to implement, test and analyze the model proposed.

In order to do this, some observations are important: The first one is that the tract length is expressed in normalized units, which can be converted into centimetres as follows

1 length unit = 0.534 cm.

The second observation is about the procedure used to "fill" the articulatory space: First, a sufficiently high number of points is uniformly generated in the hyper-rectangle defined by $\alpha_{\min}$ and $\alpha_{\max}$. After that, the corresponding log-area vectors are calculated, and those that exceed the limits defined by $\mathbf{x}_{\min}$ and $\mathbf{x}_{\max}$ are discarded, since they probably correspond either to unrealistic area functions or to areas with constrictions that are too narrow. The final observation is about the procedure used to estimate the formants associated with a given area function: They can be calculated using the wave propagation model described in [10] and the search procedure described in [3].

$$\mathbf{U}_N = \begin{bmatrix}
0.006 & -0.019 & 0.004 & -0.004 & 0.052 \\
0.010 & -0.035 & 0.026 & 0.010 & 0.130 \\
0.017 & -0.045 & 0.036 & -0.016 & 0.177 \\
0.030 & -0.026 & -0.012 & -0.110 & -0.142 \\
0.039 & -0.050 & -0.002 & -0.030 & 0.000 \\
0.057 & -0.087 & 0.023 & -0.013 & 0.136 \\
0.083 & -0.082 & 0.029 & 0.004 & 0.142 \\
0.101 & -0.071 & 0.042 & 0.034 & 0.161 \\
0.111 & -0.054 & 0.053 & 0.059 & 0.159 \\
0.113 & -0.033 & 0.059 & 0.079 & 0.137 \\
0.105 & -0.014 & 0.060 & 0.088 & 0.112 \\
0.091 & 0.001 & 0.058 & 0.090 & 0.091 \\
0.074 & 0.009 & 0.052 & 0.084 & 0.080 \\
0.072 & 0.017 & 0.048 & 0.106 & 0.102 \\
0.078 & 0.043 & 0.056 & 0.154 & 0.159 \\
0.076 & 0.059 & 0.055 & 0.170 & 0.170 \\
0.056 & 0.136 & 0.063 & 0.208 & 0.177 \\
0.011 & 0.226 & 0.092 & 0.233 & 0.199 \\
0.043 & 0.220 & 0.128 & 0.199 & 0.232 \\
0.102 & 0.205 & 0.135 & 0.131 & 0.227 \\
0.184 & 0.216 & 0.136 & 0.047 & 0.217 \\
0.264 & 0.205 & 0.130 & -0.048 & 0.192 \\
0.307 & 0.169 & 0.099 & -0.120 & 0.129 \\
0.324 & 0.125 & 0.062 & -0.157 & 0.068 \\
0.334 & 0.081 & 0.028 & -0.156 & 0.019 \\
0.340 & 0.026 & -0.010 & -0.131 & -0.020 \\
0.348 & -0.077 & -0.052 & -0.062 & -0.048 \\
0.331 & -0.208 & -0.106 & 0.112 & -0.020 \\
0.288 & -0.296 & -0.165 & 0.355 & 0.006 \\
0.202 & -0.262 & -0.247 & 0.537 & -0.002 \\
0.005 & 0.157 & -0.514 & 0.165 & 0.111 \\
0.099 & 0.304 & -0.712 & -0.259 & 0.269 \\
0.006 & -0.584 & 0.002 & -0.353 & 0.599
\end{bmatrix},$$

$$\boldsymbol{\mu}_x = \begin{bmatrix}
0.82 \\ 0.46 \\ 0.02 \\ 0.39 \\ 0.81 \\ 1.00 \\ 1.19 \\ 1.13 \\ 1.04 \\ 0.97 \\ 0.98 \\ 1.07 \\ 1.21 \\ 1.35 \\ 1.35 \\ 1.25 \\ 1.08 \\ 0.62 \\ 0.31 \\ 0.39 \\ 0.43 \\ 0.36 \\ 0.34 \\ 0.32 \\ 0.27 \\ 0.18 \\ 0.04 \\ 0.02 \\ 0.11 \\ 0.19 \\ 0.22 \\ 0.11 \\ 28.19
\end{bmatrix},$$

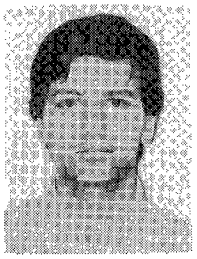
$$[\mathbf{x}_{\min}\ \mathbf{x}_{\max}] = \begin{bmatrix}
0.28 & 1.22 \\ -0.12 & 1.14 \\ -0.82 & 1.17 \\ -0.76 & 1.13 \\ -0.36 & 1.47 \\ 0.19 & 1.87 \\ 0.20 & 1.93 \\ 0.12 & 1.80 \\ -0.26 & 1.67 \\ -0.49 & 1.58 \\ -0.55 & 1.54 \\ -0.37 & 1.56 \\ 0.05 & 1.64 \\ -0.21 & 1.77 \\ -0.49 & 1.91 \\ -0.14 & 1.93 \\ -1.47 & 1.90 \\ -2.74 & 1.68 \\ -3.00 & 1.43 \\ -3.00 & 1.44 \\ -3.00 & 1.65 \\ -3.00 & 1.84 \\ -3.00 & 1.92 \\ -3.00 & 1.95 \\ -3.00 & 2.01 \\ -3.00 & 2.04 \\ -3.00 & 2.09 \\ -3.00 & 2.15 \\ -3.00 & 2.17 \\ -3.00 & 2.11 \\ -3.00 & 1.77 \\ -3.00 & 1.86 \\ 25.92 & 33.40
\end{bmatrix},$$

$$\boldsymbol{\mu}_{\alpha} = \begin{bmatrix} 0.334 \\ -1.300 \\ -0.027 \\ -0.572 \\ -0.407 \end{bmatrix}, \quad \mathbf{T}_{\alpha\beta} = \begin{bmatrix} 0.053 & -0.655 & 0.356 & 0.743 & -0.701 \\ 0.348 & 0.376 & -0.541 & -0.121 & -0.624 \\ -0.225 & -0.006 & 0.440 & -0.859 & -1.001 \\ 0.143 & 0.195 & 0.747 & 0.542 & -0.018 \\ 0.519 & -0.142 & 0.248 & -0.638 & 0.469 \end{bmatrix},$$

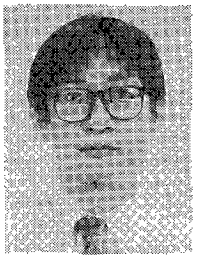
$$\boldsymbol{\alpha}_{\min}\ \boldsymbol{\alpha}_{\max} = \begin{bmatrix} -6.1201 & 8.7632 \\ -7.6430 & 3.4521 \\ -4.1455 & 4.5351 \\ -2.6781 & 2.1944 \\ -2.4002 & 2.2786 \end{bmatrix}, \quad \mathbf{U}_{\beta\gamma} = \begin{bmatrix} 0.164 & -0.458 & -0.185 & -0.854 & 0.000 \\ -0.813 & -0.030 & -0.264 & -0.083 & 0.511 \\ 0.378 & 0.046 & 0.429 & -0.045 & 0.818 \\ -0.410 & -0.012 & 0.836 & -0.253 & -0.263 \\ -0.032 & -0.887 & 0.115 & 0.444 & 0.029 \end{bmatrix},$$

$$\boldsymbol{\mu}_{\log f} = \begin{bmatrix} 2.60 \\ 3.24 \\ 3.43 \end{bmatrix}, \quad \mathbf{T}_{fg} = \begin{bmatrix} 15.4 & -7.0 & 18.5 \\ 8.3 & 23.0 & -12.7 \\ -5.2 & -7.2 & 35.4 \end{bmatrix}, \quad \mathbf{U}_{gh} = \begin{bmatrix} -0.063 & 0.970 & 0.236 \\ -0.692 & 0.128 & -0.711 \\ -0.719 & -0.208 & 0.663 \end{bmatrix},$$

$$\mathbf{M} = \begin{bmatrix} -0.457 & -0.007 & 0.068 & 0.109 & -0.810 \\ -0.105 & 0.641 & -0.399 & -0.008 & -0.126 \\ -0.084 & 0.499 & -0.140 & 0.560 & 0.237 \end{bmatrix}.$$



Hani C. Yehia was born in Belo Horizonte-MG, Brazil on November 1st, 1965. He received the degree of Engineer of Electronics in 1988 and, in 1990, completed the MSc program in the School of Electronics Engineering and Computer Science of the Instituto Tecnológico de Aeronáutica (ITA). From 1991 to 1995 he was in the Ph.D. program, being now a doctoral candidate, in the Graduate School of Engineering of Nagoya University; where he studied the relationship between acoustic and articulatory characteristics of the the human vocal-tract. He is currently with ATR-HIP; his main interest being the comprehension of the human speech communication process, as well as its applications.



Kazuya Takeda was born in Sendai, Japan on September 1, 1960. He received the B.E.E. and M.E.E. and Doctor of Engineering degrees all from Nagoya University in 1983, 1985 and 1994, respectively. In 1986, he joined ATR (Advanced Telecommunication Research Laboratories), where he was involved in the two major projects of speech database contraction and speech synthesis system development. In 1989, he moved to KDD R & D Laboratories and participated in the project for contracting a voice-activated telephone extension system. Since 1995, he has been working for Nagoya University as an Associate Professor.

Fumitada Itakura was born in Toyokawa, Japan on August 6 1940. He received the B.E.E., M.E.E. and Doctor of Engineering degrees all from Nagoya University in 1963, 1965 and 1972, respectively. In 1968, he joined the Electrical Communication Laboratory of NTT, and participated in several speech processing systems, including the maximum likelihood spectrum estimation; the PARCOR, LSP, and composite sinusoidal methods; and the APC-AB speech coding. From 1981 to 1984, he was the Head of the Speech and Acoustics Research Section of the ECL, NTT. Since 1984 he is a professor of Nagoya University, where he teaches courses of communication theory and signal processing. In 1975, he received the IEEE ASSP Senior Award for his paper on speech recognition based on the minimum prediction residual principle. He is a co-recipient with B.S. Atal of 1986 Morris N. Liebmann Award for contributions to linear predictive coding for speech processing.