Application of Time-Frequency Principal Component Analysis to Text-Independent Speaker Identification

Ivan Magrin-Chagnolleau, *Senior Member, IEEE*, Geoffrey Durou, and Frédéric Bimbot

***Abstract*—We propose a formalism, called vector filtering of spectral trajectories, that allows the integration of a number of speech parameterization approaches (cepstral analysis, Δ and $\Delta\Delta$ parameterizations, auto-regressive vector modeling, …) under a common formalism. We then propose a new filtering, called contextual principal components (CPC) or time-frequency principal components (TFPC). This filtering consists in extracting the principal components of the contextual covariance matrix, which is the covariance matrix of a sequence of vectors expanded by their context. We apply this new filtering in the framework of closed-set speaker identification, using a subset of the POLYCOST database. When using speaker-dependent TFPC filters, our results show a relative improvement of approximately 20% compared to the use of the classical cepstral coefficients augmented by their Δ-coefficients, which is significantly better with a 90% confidence level.**

***Index Terms*—Closed-set speaker identification, contextual covariance matrix, contextual principal components (CPC), POLYCOST database, speaker recognition, speech analysis, speech representation, time-frequency principal components (TFPC), vector filtering of spectral trajectories.**

I. Introduction

CEPSTRAL coefficients [1], [2] have been widely used for decades in speech processing. Although they provide a good set of feature vectors with nice properties, like a good decorrelation of the coefficients, or a good ability to deconvolve in theory the vocal source and the vocal tract filtering [2], they may not be the best solution in all situations. This is a commonly-shared point of view as proved by the abundant literature on the topic of feature parameters (see for instance [3]–[12]).

To find a good alternative to cepstral coefficients, a number of approaches have been adopted. In particular, the inability of the cepstral coefficients to extract dynamic information from speech suggested the use of Δ and $\Delta\Delta$ coefficients [13], [14]. The autoregressive (AR) vector modeling was another attempt to capture some of the dynamic information in speech signals [15]–[18].

A first aim of this paper is to propose a formalism that integrates most of these approaches under a common framework. Actually, almost every approach assumes spectral vectors as a starting point, and tries, in various ways, to extract some information from these spectral vectors by applying transformations to them. Most of these approaches can then be seen as a *vector filtering of spectral trajectories*, that is, a function applied to the coefficients of several consecutive spectral vectors [19], [20]. It is the case, for instance, of cepstral analysis [1], [2], AR vector modeling [15]–[18], Δ and $\Delta\Delta$ parameterizations [13], [14], and temporal principal components [21].

However, most of these approaches apply the filtering function only to one spectral vector, or to several of them but only component by component. We propose an other approach, based on a principal component calculation, that is applied to all the components of several consecutive spectral vectors. Such a function can be seen as a time-frequency function (or mask), that is, a function applied to all the components of a spectral vector (frequency direction) and to its time context (time direction). Since the coefficients are calculated through a principal component analysis, we call these new coefficients *contextual principal components (CPC)* or *time-frequency principal components (TFPC)* of speech [19], [20]. The latter name will be used when the original vectors contain some information about the frequency content of the signal, as spectral vectors for instance. The former name will be used otherwise.

We finally apply this new speech analysis technique in the framework of closed-set speaker identification. It leads to a significant improvement of the results compared to the classical cepstral parameterization augmented by the Δ coefficients. We obtain a relative reduction of the identification error rate of approximatively 20%.

The outline of this paper is the following. Section II illustrates the principle of the vector filtering of spectral trajectories, gives a mathematical formulation, provides an interpretation, and finally shows a couple of examples. The TFPC approach is presented in Section III, with a discussion about how to choose the components, an interpretation of the TFPC filters in terms of time-frequency masks, and a description of the use of the TFPC filtering in a pattern recognition system. An application to closed-set speaker identification is proposed in Section IV. Section V concludes this work and gives some future directions.

II. Vector Filtering of Spectral Trajectories

A. Principle of the Vector Filtering

Let $\{\mathbf{x}_t\}_{1\leq t\leq T}$ denote a sequence of spectral vectors. The principle of the vector filtering of spectral trajectories is to replace the vector $\mathbf{x}_t$ by a new vector $\mathbf{f}_t$, each coordinate of which is obtained by the application of a function to the coordinates

Manuscript received January 17, 2000; revised March 11, 2002. This work was begun while I. Magrin-Chagnolleau was with the Digital Signal Processing Group, Rice University, Houston, TX 77001 USA and finished while he was with IRISA (CNRS & INRIA), Rennes, France. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter Kabal.

I. Magrin-Chagnolleau is with CNRS, Lyon, France (e-mail: ivan@ieee.org).

G. Durou was with the Faculté Polytechnique de Mons, Mons, Belgium, when he collaborated on this work (e-mail: geoffrey.durou@debis.be).

F. Bimbot is with IRISA (CNRS & INRIA), Rennes, France (e-mail: bimbot@irisa.fr).

Digital Object Identifier 10.1109/TSA.2002.800557.

1063-6676/02$17.00 © 2002 IEEE

Authorized licensed use limited to: IEEE Xplore. Downloaded on February 17, 2009 at 23:56 from IEEE Xplore. Restrictions apply.

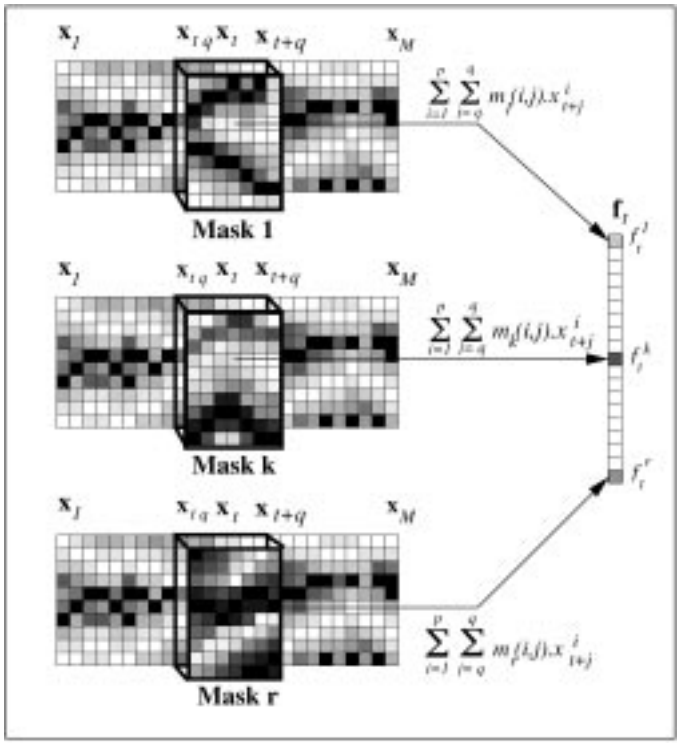

Fig. 1. Principle of the vector filtering of spectral trajectories (after [19]).

of vector $\mathbf{x}_t$ and of the preceding and the following vectors (context of vector $\mathbf{x}_t$). If the filtering is linear, this is a convolution product, which can be interpreted as the application of a time-frequency mask to a sequence of spectral vectors (see Fig. 1).

Each component of $\mathbf{f}_t$ is obtained by the application of a different function. The filtering is applied jointly in the time and the frequency directions. Each function can therefore be seen as a time-frequency mask.

This approach generalizes other filtering approaches recently presented, but that were applied only to one dimension, either the time dimension [22]–[24] or the frequency dimension [23], [25].

Our work has some similarities with the work by Milner [26]–[28]. However, we work directly on spectral coefficients instead of cepstral coefficients. This makes the interpretation of the new coefficients easier.

B. Definitions and Notations

Let $\mathbf{X}_{t-q}^{t+q}$ denote the sequence of vectors $\mathbf{x}_t$ between time $t-q$ and $t+q$

$$\mathbf{X}_{t-q}^{t+q} = \begin{bmatrix} \mathbf{x}_{t+q} \\ \vdots \\ \mathbf{x}_t \\ \vdots \\ \mathbf{x}_{t-q} \end{bmatrix}.$$

By convention, $\mathbf{x}_t = 0$ if $t \leq 0$ or $t > T$. The dimension of vector $\mathbf{X}_{t-q}^{t+q}$ is $(2q+1)\,p$.

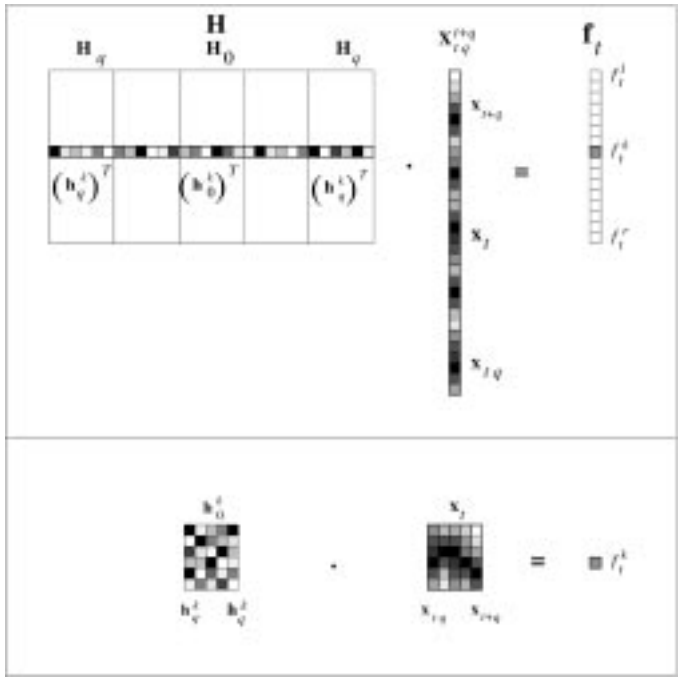

Fig. 2. Interpretation of the filtering matrix $\mathbf{H}$ in terms of time-frequency masks (after [19]).

Let $\mathcal{H}$ be a filtering operating on $\mathbf{X}_{t-q}^{t+q}$

$$\mathcal{H}\colon (\mathbb{R}^p)^{2q+1} \longrightarrow \mathbb{R}^r$$
$$\mathbf{X}_{t-q}^{t+q} \longmapsto \mathbf{f}_t = \mathcal{H}\left(\mathbf{X}_{t-q}^{t+q}\right).$$

The dimension of $\mathbf{f}_t$ is r, which can be related or not to the dimension p of the original vectors.

In the following, we only consider linear filtering. The filtering $\mathcal{H}$ can then be expressed in matrix form

$$\mathbf{H} = [\mathbf{H}_{-q} \mid \cdots \mid \mathbf{H}_0 \mid \cdots \mid \mathbf{H}_q].$$

The dimension of $\mathbf{H}$ is $r \times (2q+1)\,p$.

We also have

$$\mathbf{f}_t = \mathbf{H} \cdot \mathbf{X}_{t-q}^{t+q}$$
$$= [\mathbf{H}_{-q} \mid \cdots \mid \mathbf{H}_0 \mid \cdots \mid \mathbf{H}_q] \cdot \begin{bmatrix} \mathbf{x}_{t+q} \\ \vdots \\ \mathbf{x}_t \\ \vdots \\ \mathbf{x}_{t-q} \end{bmatrix}$$
$$= \sum_{k=-q}^{+q} \mathbf{H}_k \cdot \mathbf{x}_{t-k}.$$

The dimension of each matrix block $\mathbf{H}_k$ is $r \times p$.

C. Interpretation of the Vector Filtering

Fig. 2 proposes an interpretation of the filtering matrix $\mathbf{H}$ in terms of time-frequency masks.

The top part of the figure represents the vector filtering as it is mathematically formulated. The vector filtering is the product of a filtering matrix $\mathbf{H}$ by the expanded vector $\mathbf{X}_{t-q}^{t+q}$, and produces the filtered vector $\mathbf{f}_t$. The bottom part of the figure rep-

Authorized licensed use limited to: IEEE Xplore. Downloaded on February 17, 2009 at 23:56 from IEEE Xplore. Restrictions apply.

resents the vector filtering in terms of time-frequency masks. The vector filtering can be seen as the application of a series of time-frequency masks (a reorganization of the lines of the filtering matrix $\mathbf{H}$ by groups of p coordinates) to a sequence of spectral vectors (a reorganization of the expanded vector $\mathbf{X}_{t-q}^{t+q}$ in a sequence of consecutive spectral vectors). The dot product of the top part denotes the classical matrix product. The dot product of the bottom part is not a matrix product, but denotes a convolution between one time-frequency mask (on the figure, the kth line of the filtering matrix $\mathbf{H}$) and the sequence of spectral vectors: each element of the sequence of spectral vectors is multiplied by the corresponding element of the time-frequency mask and all the products are then added.

D. Examples of Vector Filterings

1) Cepstral Analysis: The cepstral analysis [1], [2] is one particular filtering of the spectral vectors, for which the filtering functions apply only to the frequency dimension, that is, only to $\mathbf{x}_t$. Therefore, $q = 0$. Each filtering function is a cosine transform. If the k first cosine functions are applied to $\mathbf{x}_t$, then $r = k$. And the lines of the matrix $\mathbf{H}$ are the cosine functions

$$\mathbf{H} = \begin{bmatrix} \boldsymbol{\gamma}_1^T \\ \vdots \\ \boldsymbol{\gamma}_k^T \end{bmatrix}.$$

We finally obtain

$$\mathbf{f}_t = \mathbf{H} \cdot \mathbf{X}_t^t = \begin{bmatrix} \boldsymbol{\gamma}_1^T \\ \vdots \\ \boldsymbol{\gamma}_k^T \end{bmatrix} \cdot \mathbf{x}_t = \mathbf{c}_t$$

where $\mathbf{c}_t$ is the cepstral vector corresponding to the spectral vector $\mathbf{x}_t$, and containing the cepstral coefficients 1 to k.

2) Δ and $\Delta\Delta$ Parameters: Δ and $\Delta\Delta$ parameters [13] are also examples of vector filterings. They are applied to the time dimension only. If five frames are used to calculate the Δ parameters, then $q = 2$. In fact, these parameters can be calculated for different values of q. The values $q = 1, 2, 3$ are classically chosen. If the filtered vector is composed of the original vector augmented by its Δ coefficients, then the dimension of the new vector is $r = 2p$. If the $\Delta\Delta$ coefficients are also added, then $r = 3p$. In that case, $\mathbf{H}$ can be, for instance

$$\mathbf{H} = \begin{bmatrix} \mathbf{0}_p & \mathbf{0}_p & \mathbf{I}_p & \mathbf{0}_p & \mathbf{0}_p \\ -2\mathbf{I}_p & -\mathbf{I}_p & \mathbf{0}_p & \mathbf{I}_p & 2\mathbf{I}_p \\ -\mathbf{I}_p & \mathbf{0}_p & 2\mathbf{I}_p & \mathbf{0}_p & -\mathbf{I}_p \end{bmatrix}$$

where $\mathbf{0}_p$ denotes the null matrix of dimension $p \times p$ and $\mathbf{I}_p$ the identity matrix of dimension p.

The corresponding filtered vector is

$$\mathbf{f}_t = \mathbf{H} \cdot \mathbf{X}_{t-2}^{t+2} = \begin{bmatrix} \mathbf{x}_t \\ \Delta\mathbf{x}_t \\ \Delta\Delta\mathbf{x}_t \end{bmatrix}.$$

3) Second-Order Auto-Regressive Vector Model: The application of a second-order auto-regressive model to a sequence of vectors [15]–[18] can also be interpreted as a vector filtering. $q = 2$, $r = p$, and the filtering matrix is given by

$$\mathbf{H} = [\mathbf{0}_p \mid \mathbf{0}_p \mid \mathbf{I}_p \mid \mathbf{A}_1 \mid \mathbf{A}_2]$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are the matrix coefficients of the second-order auto-regressive model. The filtered vector is the prediction error of the model at time t

$$\mathbf{f}_t = \mathbf{H} \cdot \mathbf{X}_{t-2}^{t+2} = \mathbf{x}_t + \mathbf{A}_1 \cdot \mathbf{x}_{t-1} + \mathbf{A}_2 \cdot \mathbf{x}_{t-2} = \mathbf{e}_t.$$

This filtering applies to both the time and frequency directions.

E. Composition of Filterings

Several filterings can also be composed together. In that case, the filtering matrices $\mathbf{H}$ are multiplied. For instance, Δ-cepstral parameters correspond to the composition of the cepstral filtering with the Δ filtering.

III. Time-Frequency Principal Components of Speech

We have defined a formalism that makes it possible to express a number of approaches in terms of filtering of spectral vectors. We now propose a new filtering operating in both the time and the frequency dimensions. This approach also assumes a set of training data on which to extract some principal components: it is a data-driven filtering. We call this new filtering *CPC* or *TFPC* of speech.

A. Principle of the TFPC Filtering

The idea of the TFPC filtering is to extract time-frequency patterns that are characteristic of a sequence of vectors in order to summarize the evolution of the spectral content by a few spectral sequences extracted from the entire sequence. The original sequence must be long enough to be representative of the class that we want to represent with the time-frequency patterns. This strategy can be applied to any pattern recognition problem, as long as there are enough vectors for each class to calculate the time-frequency patterns. Once the patterns have been extracted, they are used to filter the spectral vectors of both the training and the test data. And any modeling technique can then be applied to the new vectors, as it is done usually on spectral vectors or cepstral vectors, or any other vector representation of the original signal. For example, in Section IV, we apply the TFPC filtering in the framework of closed-set speaker identification, and we extract time-frequency patterns for each speaker of the training dataset.

B. Definitions and Notations

Let $\{\mathbf{x}_t\}_{1 \le t \le T}$ denote again a sequence of spectral vectors, and $\{\mathbf{x}_t^* = \mathbf{x}_t - \overline{\mathbf{x}}\}$ the sequence of the corresponding centered vectors. $\overline{\mathbf{x}}$ is the mean vector of the sequence $\{\mathbf{x}_t\}$.

Let $\boldsymbol{\mathcal{X}}_0$ denote the covariance matrix of the sequence $\{\mathbf{x}_t\}$

$$\boldsymbol{\mathcal{X}}_0 = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t - \overline{\mathbf{x}}) \cdot (\mathbf{x}_t - \overline{\mathbf{x}})^T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t^* \cdot \mathbf{x}_t^{*T}$$

and $\boldsymbol{\mathcal{X}}_k$ denote the k-lagged covariance matrix

$$\boldsymbol{\mathcal{X}}_k = \frac{1}{T}\sum_{t=k+1}^{T}(\mathbf{x}_t - \overline{\mathbf{x}}) \cdot (\mathbf{x}_{t-k} - \overline{\mathbf{x}})^T = \frac{1}{T}\sum_{t=k+1}^{T}\mathbf{x}_t^* \cdot \mathbf{x}_{t-k}^{*T}.$$

Authorized licensed use limited to: IEEE Xplore. Downloaded on February 17, 2009 at 23:56 from IEEE Xplore. Restrictions apply.

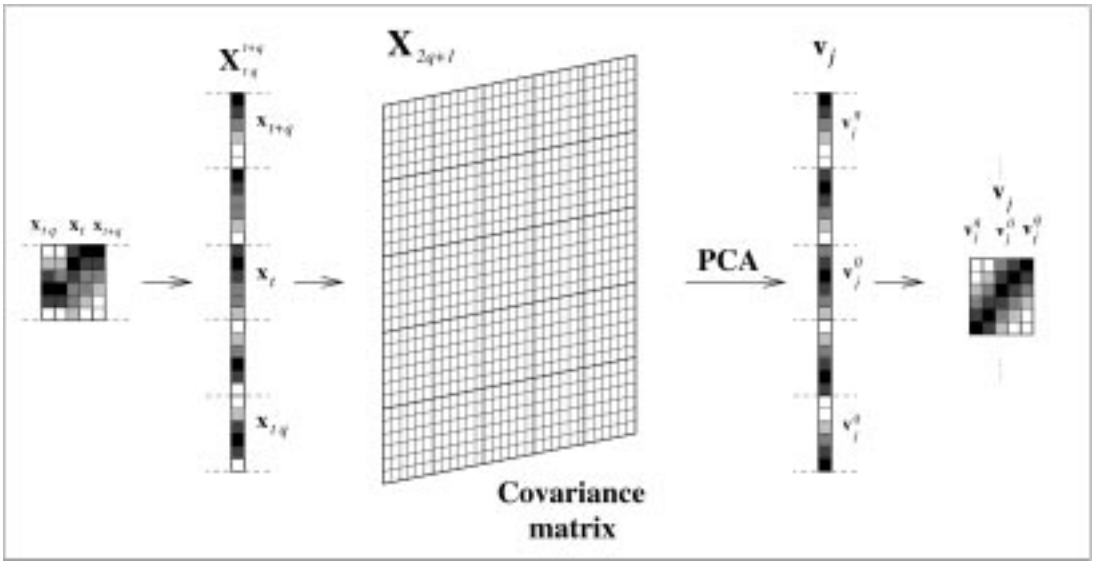

Fig. 3. Interpretation of a principal component in terms of a time-frequency mask (after [19]).

The dimension of the covariance matrix and of the lagged covariance matrices is $p \times p$.

We now define a new matrix, $\mathbf{X}_{2q+1}$, by

$$\mathbf{X}_{2q+1} = \begin{bmatrix} \boldsymbol{\mathcal{X}}_0 & \boldsymbol{\mathcal{X}}_1 & \cdots & \boldsymbol{\mathcal{X}}_{2q} \\ \boldsymbol{\mathcal{X}}_1^T & \boldsymbol{\mathcal{X}}_0 & \cdots & \boldsymbol{\mathcal{X}}_{2q-1} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\mathcal{X}}_{2q}^T & \boldsymbol{\mathcal{X}}_{2q-1}^T & \cdots & \boldsymbol{\mathcal{X}}_0 \end{bmatrix}.$$

This matrix is block-Toeplitz, and its dimension is $(2q+1)\,p \times (2q+1)\,p$.[1] This matrix can be interpreted as the covariance matrix of the vectors $\{\mathbf{X}_{t-q}^{t+q}\}_{1 \le t \le T}$, and can therefore be called *contextual covariance matrix*.

Then the principal components of this matrix [29], [30] are calculated. It is equivalent to the extraction of eigenvalues and eigenvectors of the matrix. The eigenvector associated with the largest eigenvalue is the direction of projection that preserves the maximum variance, the eigenvector associated with the second largest eigenvalue is the direction of projection that preserves the maximum variance uncorrelated (that is, orthogonal) to the first one, and so on. We finally have

$$\mathbf{X}_{2q+1} = \mathbf{V}_{2q+1} \cdot \mathbf{\Lambda}_{2q+1} \cdot \mathbf{V}_{2q+1}^T$$

with

$$\begin{aligned} \mathbf{V}_{2q+1} &= (\mathbf{v}_1, \ldots, \mathbf{v}_{2q+1}) \\ \mathbf{V}_{2q+1}^T \cdot \mathbf{V}_{2q+1} &= \mathbf{I}_{2q+1} \\ \mathbf{\Lambda}_{2q+1} &= \operatorname{diag}(\lambda_1, \ldots, \lambda_{2q+1}), \\ & \qquad \lambda_1 \ge \cdots \ge \lambda_{2q+1}. \end{aligned}$$

The dimension of matrices $\mathbf{V}_{2q+1}$ and $\mathbf{\Lambda}_{2q+1}$ is $(2q+1)p \times (2q+1)p$. The dimension of each vector $\mathbf{v}_i$, $1 \le i \le 2q+1$, is $(2q+1)p$.

[1]Although this matrix is $(2q+1)\,p \times (2q+1)\,p$, we adopt the notation $\mathbf{X}_{2q+1}$ in order to have a simpler notation. This is also the case for the matrices $\mathbf{V}_{2q+1}$ and $\mathbf{\Lambda}_{2q+1}$.

Since the principal components are extracted from the contextual covariance matrix, instead of the covariance matrix itself, we call them *CPC*. When the original vectors contain some information about the frequency content of the signal, as spectral vectors, we can call these components more specifically *TFPC*.

C. Choice of the Components

Once the principal components have been calculated, it is necessary to select which ones to keep. It is common to keep the first components, their number depending on the experiment. Since the eigenvalues correspond to a variance measurement, a criterion for keeping them can be a percentage of the total variance, for instance 80% [29], [30]. Some other procedures can be used for the choice of the components like the F-ratio [31], [32] or the knock-out procedure [33]. Here is an example of the filtering matrix $\mathbf{H}$, corresponding to the selection of the first eight components

$$\mathbf{H} = [\mathbf{v}_1, \ldots, \mathbf{v}_8]^T.$$

If all the components are kept, then $\mathbf{H} = \mathbf{V}_{2q+1}^T$.

D. Interpretation

Each principal component can be interpreted in terms of a time-frequency mask, as shown in Fig. 3.

On the left part of the figure, a sequence of spectral vectors $\{\mathbf{x}_t\}_{1 \le t \le T}$ is represented. Then, the corresponding expanded vectors $\{\mathbf{X}_{t-q}^{t+q}\}_{1 \le t \le T}$ are formed. The covariance matrix of these expanded vectors, $\mathbf{X}_{2q+1}$, is calculated. It is a block-Toeplitz matrix. Then the principal component analysis is applied and the eigenvectors of this matrix are obtained. The dimension of each eigenvector $\mathbf{v}_i$, $1 \le i \le 2q+1$, is $(2q+1)p$. Each of them can be reordered as a sequence of p-dimensional vectors to form a time-frequency mask. Finally, for each eigenvector, a time-frequency mask is obtained, as shown on the right part of the figure.

Authorized licensed use limited to: IEEE Xplore. Downloaded on February 17, 2009 at 23:56 from IEEE Xplore. Restrictions apply.

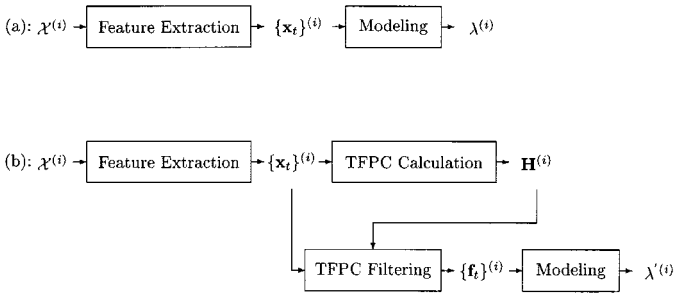

Fig. 4. Training phase of a pattern recognition system: (a) without the TFPC filtering and (b) with the TFPC filtering.

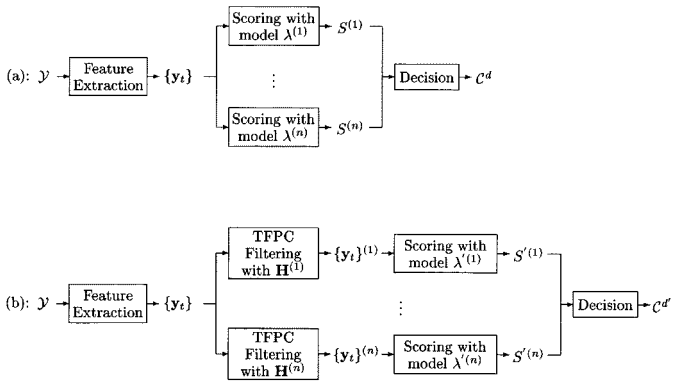

Fig. 5. Test phase of a pattern recognition system: (a) without the TFPC filtering and (b) with the TFPC filtering.

E. TFPC Filtering and Pattern Recognition

The next paragraphs explain how to integrate the TFPC filtering in a speaker recognition system.[2]

1) Training Phase: Fig. 4 shows the training phase of the recognition system: (a) without the TFPC filtering; and (b) with the TFPC filtering. In both cases, we dispose of a training pattern $\mathcal{X}^{(i)}$ representing the class $\mathcal{C}^{(i)}$. Feature parameters are extracted from this training pattern (for instance spectral vectors) and the sequence of vectors $\{\mathbf{x}_t\}^{(i)}$ is formed.

In the first case, this sequence of vectors is then modeled (for instance using a Gaussian mixture model [34]–[36]) and a model $\lambda^{(i)}$ is obtained for class $\mathcal{C}^{(i)}$. At the end, the class $\mathcal{C}^{(i)}$ is represented by the model $\lambda^{(i)}$.

In the second case, some time-frequency principal components are first extracted from the feature vectors $\{\mathbf{x}_t\}^{(i)}$ and a filtering matrix $\mathbf{H}^{(i)}$ is obtained for the class $\mathcal{C}^{(i)}$. This matrix is then used to filter the sequence $\{\mathbf{x}_t\}^{(i)}$ and a sequence of new vectors $\{\mathbf{f}_t\}^{(i)}$ is formed. Finally, this new sequence of vectors is modeled and a model $\lambda'^{(i)}$ is obtained for class $\mathcal{C}^{(i)}$. At the end, the class $\mathcal{C}^{(i)}$ is represented by a model $\lambda'^{(i)}$ and a filtering matrix $\mathbf{H}^{(i)}$.

2) Test Phase: Fig. 5 represents the test phase of the recognition system: (a) without the TFPC filtering and (b) with the TFPC filtering. In both cases, the test pattern $\mathcal{Y}$ has to be classified in one of the reference classes $\mathcal{C}^1$ to $\mathcal{C}^n$, where n is the number of classes. First, parameter vectors are extracted from the pattern $\mathcal{Y}$, and the sequence $\{\mathbf{y}_t\}$ is formed.

[2]This approach can easily be reused for other pattern recognition problems.

In the first case, a score is calculated on this sequence with each reference model, $\lambda^{(1)}$ to $\lambda^{(n)}$, and the n scores S^1 to S^n are obtained. Then, a decision algorithm is applied to decide to which class $\mathcal{C}^d$ the test pattern $\mathcal{Y}$ belongs.

In the second case, the TFPC filtering of each reference class, $\mathbf{H}^{(1)}$ to $\mathbf{H}^{(n)}$, is applied to the sequence $\{\mathbf{y}_t\}$. This leads to the new sequences $\{\mathbf{y}_t\}^{(1)}$ to $\{\mathbf{y}_t\}^{(n)}$. Then, a score is calculated on these sequences with each corresponding reference model, $\lambda'^{(1)}$ to $\lambda'^{(n)}$, and the n new scores S'^1 to S'^n are obtained. Finally, a decision algorithm is applied to decide to which class $\mathcal{C}'^d$ the test pattern $\mathcal{Y}$ belongs.

IV. Application of the TFPC Filtering to Closed-Set Speaker Identification

We have defined the TFPC filtering and explained how to integrate it in a speaker recognition system. To evaluate the effectiveness of this approach, we now apply it to closed-set speaker identification [20].

A. Experimental Protocol

1) Task: The task considered in our experiments is closed-set text-independent speaker identification, which consists in recognizing the identity of the speaker of a test utterance, without using any phonetic information about the utterance. Closed-set means that the system always gives the identity of one of the speakers of the reference set as the answer.

2) Database: In our experiments, a subset of the POLYCOST database [37] was used. This subset is composed of 112 speakers (64 females and 48 males), who have pronounced some free-text utterances and some guided utterances in English through the telephone. Speakers were originating from various European countries, and most of them were not English native speakers. Free-text utterances were used for training, guided utterances were used for testing. Both the training and the test data were recorded during 10 sessions in average, spread over three months. We used about 90 s of speech for the training, and several utterances of 5 s in average for the test. The total number of tests was 560.

3) Spectral Analysis: Each utterance was analyzed as follows: the speech signal was decomposed in frames of 30 ms at a frame rate of 10 ms. A Hamming window was applied to each frame. The signal was pre-emphasized with a coefficient equal to 0.95. For each frame, a fast Fourier transform was computed and provided 256 square module values representing the short term power spectrum in the 0–4 kHz band. This Fourier power spectrum was then used to compute 13 filter bank coefficients. Each filter was triangular. The filters were placed on a nonuniform frequency scale, similar to the Bark/Mel scale. The central frequency of the 13 filters were, in Hz: 102, 219, 353, 506, 682, 883, 1114, 1378, 1681, 2028, 2425, 2881, and 3402. Each filter covered a spectral range from the central frequency of the previous filter to the central frequency of the next filter, with a maximum value of one for its own central frequency. We finally took the base 10 logarithm of each filter output and multiplied the result by 10, to form a 13-dimensional vector of filter bank coefficients in decibels.

Authorized licensed use limited to: IEEE Xplore. Downloaded on February 17, 2009 at 23:56 from IEEE Xplore. Restrictions apply.

4) Application of the TFPC Filtering to Spectral Vectors: We then calculated for each speaker the TFPC for the values $q = 0, 1, 2$, which corresponds to one frame (the classical principal components analysis), three frames, and five frames, respectively. For each value of q, all the components were kept. Then, the extracted speaker-dependent TFPC filtering was applied to the initial spectral vectors in order to obtain the new vectors. The filtering matrices were kept to be used also during the test phase. As an additional experiment, TFPC were also extracted from all the training data pooled together (see [19]). A speaker-independent TFPC filtering, calculated on multispeaker speech, was obtained. Various values of q (0, 1, and 2) were also tested, and all the components were kept for each value of q. The last filtering that was tested was a classical Δ calculation [13], [14], using three or five frames, which provided us with 26-dimensional parameter vectors.

5) Baseline: Cepstrum and Δ-Cepstrum: For comparison, cepstral coefficients [1], [2] were also extracted from the spectral vectors. The first coefficient c_0, which is highly correlated with the energy of the frame, was not kept. Only c_1 to c_{12} were used. The corresponding Δ parameters [13], [14] were calculated using three or five frames, and appended to the cepstral coefficients. Finally, 24-dimensional parameter vectors were obtained. To complete our experiments, both speaker-dependent and speaker-independent TFPC filterings were also tested on the cepstral vectors, which is equivalent to composing a cepstral filtering and a TFPC filtering, and applying the composition directly to the spectral vectors.

6) Modeling: A Gaussian mixture model (GMM) [34]–[36] was used to represent each reference speaker. Each mixture was composed of eight components, and diagonal covariance matrices were chosen. The Gaussian mixture models were trained using the expectation-maximization (EM) algorithm [34], [35], [38]. The initialization of the EM algorithm was done with a vector quantization (VQ) algorithm [39].

B. Results and Discussion

Identification error rates are reported in Table I. We also give, in the discussion, the 95% and the 90% confidence intervals of the baseline score assuming that the probability distribution of the error rate is binomial. If R is the identification error rate of one experiment and N the total number of tests, then the confidence interval is given by [30], [40]

$$R \pm u \cdot \sqrt{\frac{R(1-R)}{N}}$$

with $u = 1.96$ for the 95% confidence interval and $u = 1.65$ for the 90% confidence interval.

The top part of Table I presents classical results using spectral or cepstral coefficients and Δ parameters. The baseline score, 11.43%, was obtained using cepstral coefficients augmented by the Δ parameters, with a Δ calculation over five vectors. Cepstral coefficients perform better than spectral coefficients when using the Δ parameters, which is a very classical result in speaker identification.

TABLE I
SPEAKER IDENTIFICATION RESULTS WHEN USING THE Δ PARAMETERS, ONE TFPC FILTER FOR ALL SPEAKERS, OR ONE TFPC FILTER FOR EACH SPEAKER. SEVERAL SIZES OF CONTEXT (1, 3, AND 5 VECTORS) WERE TESTED. RESULTS ARE GIVEN AS A PERCENTAGE OF IDENTIFICATION ERROR

Coefficients	static	static + Δ	
Context	1 vector	3 vectors	5 vectors
Spectral	21.96	19.64	18.93
Cepstral	15.71	12.32	**11.43**

Coefficients	Speaker-independent TFPC filtering		
Context	1 vector	3 vectors	5 vectors
Spectral	11.96	**11.35**	11.61
Cepstral	11.43	12.50	16.25

Coefficients	Speaker-dependent TFPC filtering		
Context	1 vector	3 vectors	5 vectors
Spectral	9.82	**9.11**	9.82
Cepstral	10.00	9.29	10.71

The middle part of Table I presents results using a speaker-independent TFPC filtering. When cepstral vectors were used as original vectors, error rates were not better than the baseline score. The error rates even degrade if we use three or five vectors. When spectral vectors were used as original vectors, the best error rate obtained was 11.35%, which is not significantly better than the baseline score given its 95% confidence interval (8.79%, 14.07%).

More interesting results were obtained using speaker-dependent TFPC filterings as reported in the bottom part of Table I. The best error rate, 9.11%, outperformed the baseline score of about 20% relatively. The figure of 9.11% falls near the border of the 95% confidence interval of the baseline score (8.79%, 14.07%). However, it falls outside the 90% confidence interval (9.21%, 13.65%). Therefore, the difference observed is significant with a 90% confidence level.

V. CONCLUSIONS AND PERSPECTIVES

In this paper, we have presented a new formalism, called *vector filtering of spectral trajectories*, that integrates under a common formalism several approaches such as cepstral analysis, Δ and $\Delta\Delta$ parameters, or auto-regressive vector models. In the framework of this formalism, we proposed a new filtering called *time-frequency principal components (TFPC)* or, more generally, *contextual principal components (CPC)*, which consists of calculating the principal components of the contextual covariance matrix, i.e., the covariance matrix of a sequence of vectors augmented by their time context. We also explained how to integrate the new filtering in a speaker recognition system. We applied the new filtering to closed-set speaker identification, with several sizes of context, and using spectral or cepstral vectors as original vectors. Our best configuration, based on speaker-dependent TFPC filterings, outperforms the classical parameterization using cepstral and Δ-cepstral coefficients by approximatively 20%, which is significantly better with a 90% confidence level.

A number of issues are still pending. One of them is the selection of the components. In our experiments, we systematically kept all the components. An interesting experiment would

Authorized licensed use limited to: IEEE Xplore. Downloaded on February 17, 2009 at 23:56 from IEEE Xplore. Restrictions apply.

be the removal of the last components, which should not decrease noticeably the performance, and would provide a more compact representation of the signal. However, in the case of speaker-dependent TFPC filterings, the choice of the components to keep must be done more carefully, as some parameters having a small variance may still be a very good indicator of a particular speaker.

Another extension of this work is the application of the TFPC filtering to speaker verification and speaker tracking rather than speaker identification. However, in these cases, a score normalization is needed, and the most common approach is the calculation of a log-likelihood ratio using a nonspeaker model. Therefore, a theoretical study of the normalization problem needs to be done.

The TFPC filtering can also be applied to speech recognition. It may be useful to neutralize some of the effect of the speaker variability and/or the transmission channel variability.

More generally, the TFPC approach can be investigated on for a number of pattern recognition problems, and we hope that this preliminary work will encourage some of our colleagues to experiment with this approach.

REFERENCES

[1] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, M. Rosenblatt, Ed. New York: Wiley, 1963, ch. 15, pp. 209–243.

[2] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Tran. Audio Electroacoust.*, vol. AE-16, pp. 221–226, June 1968.

[3] L. Xu, "Perceptually-based features for speaker identification," Ph.D. dissertation, Univ. College Swansea, Univ. Wales, Wales, U.K., Feb. 1992.

[4] Z. P. Sun and J. S. Mason, "Combining features via LDA in speaker recognition," in *Proc. EUROSPEECH 93*, vol. 3, Berlin, Germany, Sept. 1993, pp. 2287–2290.

[5] J. Thompson and J. S. Mason, "Within class optimization of cepstra for speaker recognition," in *Proc. EUROSPEECH 93*, vol. 1, Berlin, Germany, Sept. 1993, pp. 165–168.

[6] X. Wang and G. Zhao, "Text-dependent speaker verification using recurrent time delay neural networks for feature extraction," in *Proc. ICSP 93*, vol. 1, Beijing, China, Oct. 1993, pp. 674–677.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[8] N. Higuchi and M. Hashimoto, "Analysis of acoustic features affecting speaker identification," in *Proc. EUROSPEECH 95*, vol. 1, 1995, pp. 435–438.

[9] C. R. Jankowski Jr., T. F. Quatieri, and D. A. Reynolds, "Fine structure features for speaker identification," in *Proc. ICASSP 96*, vol. 2, Atlanta, GA, May 1996, pp. 689–692.

[10] T. Kitamura and S. Takei, "Speaker recognition model using two-dimensional mel-cepstrum and predictive neural network," in *Proc. ICSLP 96*, 1996.

[11] C.-S. Liu, "A general framework of feature extraction: Application to speaker recognition," in *Proc. ICASSP 96*, vol. 2, Atlanta, GA, May 1996, pp. 669–672.

[12] L. Besacier and J.-F. Bonastre, "Subband approach for automatic speaker recognition: Optimal division of the frequency domain," in *Proc. Workshop Audio and Video Biometric Person Authentification*, Craus-Montana, Switzerland, 1997, pp. 195–202.

[13] S. Furui, "Comparison of speaker recognition methods using static features and dynamic features," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 3, pp. 342–350, June 1981.

[14] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 871–879, June 1988.

[15] Y. Grenier, "Utilization de la prédiction linéaire en reconnaissance et adaptation au locuteur," in *XIèmes Journées d'Etude sur la Parole*, Strasbourg, France, May 1980, pp. 163–171.

[16] F. Bimbot, L. Mathan, A. de Lima, and G. Chollet, "Standard and target-driven AR-vector models for speech analysis and speaker recognition," in *Proc. ICASSP 92*, vol. 2, San Francisco, CA, Mar. 1992, pp. II.5–II.8.

[17] C. Montacié, P. Deléglise, F. Bimbot, and M.-J. Caraty, "Cinematic techniques for speech processing: Temporal decomposition and multivariate linear prediction," in *Proc. ICASSP 92*, vol. 1, San Francisco, CA, Mar. 1992, pp. 153–156.

[18] C. Montacié and J.-L. Le Floch, "AR-vector models for free-text speaker recognition," in *Proc. ICSLP 92*, vol. 1, Banff, AB, Canada, Oct. 1992, pp. 611–614.

[19] I. Magrin-Chagnolleau, "Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte," Ph.D. dissertation, École Nat. Supérieure Télécommun., Jan. 1997.

[20] I. Magrin-Chagnolleau and G. Durou, "Time-frequency principal components of speech: Application to speaker identification," in *Proc. EUROSPEECH 99*, Budapest, Hungary, Sept. 1999, pp. 759–762.

[21] F. Bimbot, E. Bocchieri, and B. Atal, "Sous-espaces de projection de séquences de trames acoustiques pour l'analyze et la reconnaissance de parole," in *XXIèmes Journées d'Etude sur la Parole*, Avignon, France, 1996.

[22] J.-L. Shen, W.-L. Hwang, and L.-S. Lee, "Robust speech recognition features based on temporal trajectory filtering of frequency band spectrum," in *Proc. ICSLP 96*, 1996.

[23] C. Nadeu, J. B. Marino, J. Hernando, and A. Nogueiras, "Frequency and time filtering of filter-bank energies for HMM speech recognition," in *Proc. ICSLP 96*, 1996.

[24] J. P. Openshaw and J. S. Mason, "Noise robust estimate of speech dynamics for speaker recognition," in *Proc. ICSLP 96*, 1996.

[25] D. J. Darlington and D. R. Campbell, "Sub-band adaptive filtering applied to speech enhancement," in *Proc. ICSLP 96*, 1996.

[26] B. P. Milner and S. V. Vaseghi, "An analysis of cepstral-time matrices for noise and channel robust speech recognition," in *Proc. EUROSPEECH 95*, vol. 1, 1995, pp. 519–522.

[27] B. Milner, "Inclusion of temporal information into features for speech recognition," in *Proc. ICSLP 96*, 1996.

[28] N. Harte, S. Vaseghi, and B. Milner, "Dynamic features for segmental speech recognition," in *Proc. ICSLP 96*, 1996.

[29] I. T. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer-Verlag, 1986.

[30] G. Saporta, *Probabilités, analyse des données et statistique*. Paris, France: Éditions Technip, 1990.

[31] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, pt. 2, vol. 51, no. 6, pp. 2044–2056, 1972.

[32] S. M. Kendall and A. Stuart, *The Advanced Theory of Statistics*. London, U.K.: Griffin, 1977.

[33] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 176–182, Apr. 1975.

[34] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[35] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. London, U.K.: Marcel Dekker, 1988.

[36] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 72–83, Jan. 1995.

[37] (1998) *Speaker Recognition in Telephony*. Eur. COST 250 Action. [Online]. Available: http://circhp.epfl.ch/polycost.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 6, no. 39, pp. 1–38, 1977.

[39] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantization design," *IEEE Trans. Commun.*, vol. 28, pp. 84–95, Jan. 1980.

[40] T. H. Wonnacott and R. J. Wonnacott, *Introductory Statistics for Business and Economics*. New York: Wiley, 1990.

Authorized licensed use limited to: IEEE Xplore. Downloaded on February 17, 2009 at 23:56 from IEEE Xplore. Restrictions apply.

Ivan Magrin-Chagnolleau (S'94–M'97–SM'02) was born in Paris, France, in 1969. He received the engineer diploma in electrical engineering from the ENSEA, Cergy-Pontoise, France, in 1992, the M.S. degree in electrical engineering from Paris XI University, Orsay, France, in 1993, the M.A. degree in phonetics from Paris III University in 1996, and the Ph.D. degree in electrical engineering from the ENST, Paris, in 1997.

In February 1997, he joined the Speech and Image Processing Services Laboratory, AT&T Labs Research, Florham Park, NJ. In October 1998, he visited the Digital Signal Processing Group, Electrical and Computer Engineering Department, Rice University, Houston, TX. In October 1999, he joined IRISA, Rennes, France. From October 2000 to August 2001, he was an Assistant Professor with the Computer Science Laboratory, University of Avignon, Avignon, France. In October 2001, he became a permanent Researcher with the French National Center for Scientific Research and is currently working at the "Laboratoire Dynamique Du Langage," one of the CNRS associated laboratories. His research interests include audio indexing, speaker recognition, language identification, speech recognition, statistical pattern recognition, signal representations and decompositions, and data analysis.

Dr. Magrin-Chagnolleau is a member of the IEEE Signal Processing Society, the IEEE Computer Society, and the International Speech Communication Association (ISCA).

Geoffrey Durou was born in Tournai, Belgium, in 1973. He received the engineer diploma from the Faculté Polytechnique de Mons (FPM), Mons, Belgium, in June 1996.

In September 1996, he joined the Speech and Image Processing Laboratory, FPM. His research interests include vector quantization, speaker recognition, signal representations and decompositions.

Frédéric Bimbot was born in Neuilly-sur-Seine, France, in 1963. He received the telecommunication engineer degree and the Ph.D. degree in signal processing on the topic of speech synthesis using temporal decomposition from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1985 and 1988, respectively. He also received the B.A. degree in linguistics from the Sorbonne Nouvelle University, Paris III, in 1987.

In 1990, he joined CNRS (the French National Center for Scientific Research) and became a permanent Researcher. He continued working for ENST (one of the CNRS-associated laboratories) for 7 years and then moved to IRISA (another CNRS-associated laboratory, also part of INRIA), where he has been working since 1997. He visited AT&T Bell Laboratories several times in the past few years (for a total of 15 months). He has been involved in many national and European projects, in particular Esprit projects SPRINT on speech recognition using neural networks (1989–1990), SAM-A on assessment methodology (1992–1993), and DiVAN on audio data indexing (1997–2000). He was also the Research Work-Package Manager in the Telematics projects CAVE (1995–1997) and PICASSO (1998–2000), both on speaker verification. His research interests are focused on speech and audio signal analysis and modeling, speaker characterization and verification, speech recognition and assessment methodology of speech systems. Since 2001, he is heading at IRISA a research group called METISS, centered on these topics.

Dr. Bimbot was the Chairman of the GFCP (a branch of the French Acoustic Society dedicated to spoken communication) from 1996 to 2000 and since 1998, he has been a member of the ISCA Board (International Speech Communication Association, formerly known as ESCA).

Authorized licensed use limited to: IEEE Xplore. Downloaded on February 17, 2009 at 23:56 from IEEE Xplore. Restrictions apply.