TIME-FREQUENCY PRINCIPAL COMPONENTS OF SPEECH: APPLICATION TO SPEAKER IDENTIFICATION

Ivan Magrin-Chagnolleau and Geoffrey Durou***

*Rice University, Houston, Texas, USA - **Faculté Polytechnique de Mons, Mons, Belgium
ivan@ieee.org - durou@tcts.fpms.ac.be

ABSTRACT

In this paper, we propose a formalism, called vector filtering of spectral trajectories, which allows to integrate under a common formalism a lot of speech parameterization approaches. We then propose a new filtering in this framework, called time-frequency principal components (TFPC) of speech. We apply this new filtering in the framework of speaker identification, using a subset of the POLYCOST database. The results show an improvement of roughly 20 % compared to the use of the classical cepstral coefficients augmented by their Δ coefficients.

1. INTRODUCTION

Cepstral coefficients [8] 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 their ability to decorrelate in theory the vocal source and the vocal tract filtering [8], we are convinced that they are not the ultimate solution to represent speech signals in most of the situations.

To find a good alternative to cepstral coefficients, a lot of approaches have been adopted. In particular, the lack of ability of the cepstral coefficients to extract dynamic information from speech suggested the use of Δ and $\Delta\Delta$ coefficients [2]. The Auto-Regressive (AR) vector modeling was another attempt to capture dynamic information of speech [3].

A first aim of this paper is to integrate most of these approaches under a common formalism. Actually, almost every approach assumes spectral vectors as a starting point, and tries, in different ways, to extract some information from these spectral vectors by applying different transformations on 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.

However, most of the previous approaches apply the filtering function only to one spectral vector, or to several of them but only component by component. We propose a new filtering, based on a principal component calculation, which 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 *time-frequency principal components (TFPC)* of speech. We also show how the application of this filtering is an attempt to capture dynamic information from speech. We finally apply this new speech analysis in the framework of speaker identification, which allow us to improve considerably the results compared to the classical cepstral parameterization augmented by the Δ coefficients.

2. VECTOR FILTERING OF SPECTRAL TRAJECTORIES

2.1. Principle

Let $\{\mathbf{x}_t\}_{1\leq t\leq M}$ 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 $\mathbf{f}_t$, whose each component is obtained by the application of a function to the coordinates of vector $\mathbf{x}_t$ and of the preceeding and following vectors (context of vector $\mathbf{x}_t$). This is a convolution product, which can be interpreted as the application of a time-frequency mask to a sequence of spectral vectors (see Figure 1). We can see on this figure that each component of $\mathbf{f}_t$ is obtained by the application of a different function. We can also see that the filtering is applied jointly in the time and the frequency directions. Each function can thus be seen as a time-frequency mask.

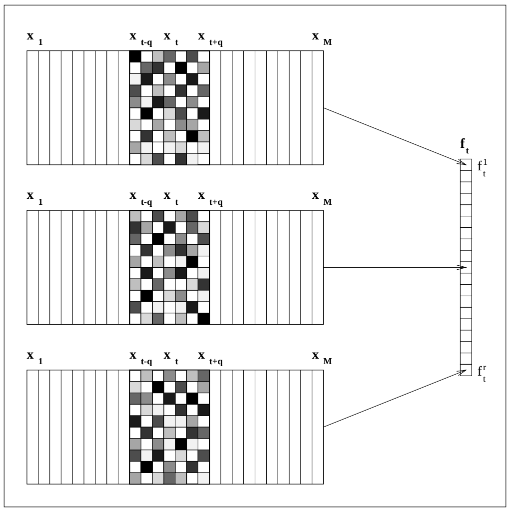

Figure 1: *Principle of the vector filtering of spectral trajectories (after [5]).*

The approach presented by Milner [7] is quite similar. However, he adopts cepstral coefficients as a starting point, whereas we work directly on spectral coefficients, which eases the interpretation of the new coefficients, and also simplifies the whole process. He also applies the filtering only to the time dimension, and not to the frequency dimension.

Figure 2 illustrates several classical approaches in term of vector filtering. The first example (a) represents the application of a first derivative approximation function to the first coordinate of several consecutive vectors, which corresponds to the calculation of a Δ coefficient on the first coordinate. The second example (b) is also a Δ coefficient but calculated on the second coordinate. The third example (c) shows the application of a cosine transform to a spectral vector, which corresponds to the calculation of a cepstral coefficient. The last example (d) shows the application of a cosine transform to the first coordinate of several consecutive vectors.

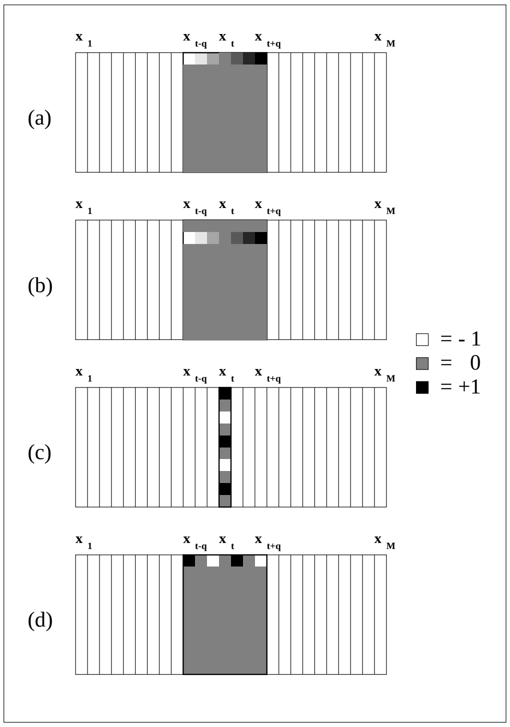

Figure 2: *Examples of vector filtering of spectral trajectories (after [5]): (a) Δ parameter on the first coordinate; (b) Δ parameter on the second coordinate; (c) cepstral coefficient; (d) cosine transform of the first coordinate using 7 consecutive vectors.*

Once again, we can notice that all these examples show a filtering which operates only in one direction, either the time one or the frequency one, whereas the general approach, as illustrated in Figure 1, operates jointly in both directions.

2.2. Definitions and Notation

Let $\{\mathbf{x}_t^* = \mathbf{x}_t - \bar{\mathbf{x}}\}_{1 \leq t \leq M}$ denote the sequence of centered vectors corresponding to the sequence $\{\mathbf{x}_t\}_{1 \leq t \leq M}$, where $\bar{\mathbf{x}}$ is the corresponding mean vector:

$$\bar{\mathbf{x}} = \frac{1}{M} \sum_{t=1}^{M} \mathbf{x}_t$$

We also define 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 $t > M$. The dimension of vector $\mathbf{X}_{t-q}^{t+q}$ is $(2q+1)\,p$.

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

$$\begin{aligned} \mathcal{H} : (\mathbb{R}^p)^{2q+1} &\longrightarrow \mathbb{R}^r \\ \mathbf{X}_{t-q}^{t+q} &\longmapsto \mathbf{f}_t = \mathcal{H}(\mathbf{X}_{t-q}^{t+q}) \end{aligned}$$

The dimension of $\mathbf{f}_t$ is r.

In the following, we only consider the case of a linear filtering. In that case, the filtering $\mathcal{H}$ can be expressed in a matricial form:

$$\mathbf{H} = [\, \mathbf{H}_{-q} \mid \ldots \mid \mathbf{H}_0 \mid \ldots \mid \mathbf{H}_q \,]$$

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

And we have:

$$\begin{aligned} \mathbf{f}_t &= \mathbf{H} \cdot \mathbf{X}_{t-q}^{t+q} \\ &= \sum_{k=-q}^{+q} \mathbf{H}_k \cdot \mathbf{x}_{t-k}^* \end{aligned}$$

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

2.3. Examples of Vector Filterings

2.3.1. Cepstral Analysis

The cepstral analysis [8] is one particular filtering of the spectral vectors, where the filtering functions are applied only to the frequency dimension, that is, only to $\mathbf{x}_t^*$, thus $q = 0$. Each filtering function is a cosine transform. If we apply only the k first cosine transforms to $\mathbf{x}_t^*$, then $r = k$. And the lines of the matrix $\mathbf{H}$ are the different cosine functions:

$$\mathbf{H} = \begin{bmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_k^T \end{bmatrix}$$

We finally obtain:

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

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

2.3.2. Δ and $\Delta\Delta$ Parameters

Δ and $\Delta\Delta$ parameters [2] are also examples of vector filterings. They are applied to the time dimension only. If we use 5 frames to calculate the Δ parameters, then $q = 2$. In fact, these parameters can be calculated for different values of q. We classically use the values $q = 1, 2, 3$. 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 we also add the $\Delta\Delta$ coefficients, 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:

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

2.3.3. *Second-Order Auto-Regressive Vector Model*

The application of a second-order auto-regressive model to a sequence of vectors [3] can also be interpreted as a vector filtering. In that case, we have $q = 2$, $r = p$, 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 matricial 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 apply to both the time and frequency directions.

2.4. Composition of Filtering

These different filterings can also be composed together, that is, the matrices $\mathbf{H}$ can be multiplied. This is the case for instance when we compute the Δ cepstral parameters. We compose the cepstral filtering with the Δ filtering.

3. TIME-FREQUENCY PRINCIPAL COMPONENTS OF SPEECH

We have defined a formalism which allows to express a lot of approaches in terms of filtering of spectral vectors. However, as it has been seen, most of these classical approaches do not combine the time and the frequency dimensions in the same filtering (except for the second-order AR-vector model). We propose a new filtering operating in both dimensions. This new filtering also assumes a set of training data on which to extract some principal components, which means that it is a data-driven filtering. We call this new filtering *Time-Frequency Principal Components (TFPC)* of Speech.

3.1. Principle of the TFPC Filtering

The idea of the TFPC filtering is to extract time-frequency patterns which are characteristic of the sequence of training vectors, that is, to summarize the evolution of the spectral content by a few spectral sequences extracted from the entire sequence. The original sequence has thus to be long enough, and representative of the class we want to represent with the time-frequency patterns. This strategy can be applied to any pattern recognition problem, as long as we have 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 datasets. And any modeling technique can then be applied on the new vectors, as it is done usually on spectral vectors or cepstral vectors, or any other vector representation of the original signal. As an example, we apply the TFPC filtering in the framework of closed-set text-independent speaker identification, and we extract time-frequency patterns for each speaker of the training database.

3.2. Definition and notation

Let $\{\mathbf{x}_t\}_{1 \leq t \leq M}$ denote again a sequence of spectral vectors, and $\{\mathbf{x}_t^*\}$ the sequence of the corresponding centered vectors.

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

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

and $\boldsymbol{\mathcal{X}}_k$ the lagged covariance matrix at the order k:

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

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:

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

Note that this matrix is block-Toeplitz, and that its dimension is $(2q+1)\ p \times (2q+1)\ p$. This matrix can be interpreted as the covariance matrix of the vectors $\{\mathbf{X}_{t-q}^{t+q}\}_{1 \leq t \leq M}$.

We now calculate the principal components of this matrix [4]. It is equivalent to the extraction of eigenvalues and eigenvectors of the matrix. The eigenvector associated with the largest eigenvalue is then the direction of projection which conserves the maximum of the variance; The eigenvector associated to the second largest eigenvalue is the direction of projection which conserves the maximum of the variance uncorrelated (that is orthogonal) to the first one; And so on. We have then:

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

with:

$$\begin{aligned} \mathbf{V}_{2q+1} &= (\mathbf{v}_1, \dots, \mathbf{v}_{2q+1}) \\ \boldsymbol{\Lambda}_{2q+1} &= diag(\lambda_1, \dots, \lambda_{2q+1}), \quad \lambda_1 \geq \dots \geq \lambda_{2q+1} \end{aligned}$$

The dimension of the matrix $\mathbf{V}_{2q+1}$ and of the matrix $\mathbf{M}_{2q+1}$ is $(2q+1)p \times (2q+1)p$. The dimension of each vector $\mathbf{v}_i$, $1 \leq i \leq 2q+1$, is $(2q+1)p$.

3.3. Choice of the Components

Once the principal components have been calculated, we have to decide which ones to keep. Since the decomposition is done according to the maximum of the variance, the first components contain a lot of information, and the last components correspond mainly to noise. The first components are usually kept, the number of them depending on the experiment. Since the eigenvalues correspond to a variance measurement, the criterion for keeping them can be a percentage of the total variance, for instance 80 % [4]. Some other procedures can be used for the choice of the components like the F-ratio or the knock-out procedure. Whatever the method is, the choice of the components can only be done experimentally.

4. APPLICATION TO SPEAKER IDENTIFICATION

4.1. Task

We have tested the TFPC of speech in the framework of closed-set text-independent speaker identification. There is a single reference per speaker. The possibility of rejection is not taken into account: the test speaker is always part of the set of references.

4.2. Database

We use a subset of the POLYCOST database [1], a telephone database, containing 112 speakers (64 females and 48 males). For each speaker, we use in average 90 seconds of speech for the training, and several utterances of 5 seconds in average for the tests. The total number of tests is 560.

4.3. Spectral Analysis

Each utterance is analyzed as followed : the speech signal is decomposed in frames of 30 ms at a frame rate of 10 ms. A Hamming window is applied to each frame. The signal is pre-emphasized with a coefficient 0.95. For each frame, a fast Fourier transform is computed and provides 252 square module values representing the short term power spectrum in the 0-4 kHz band. This Fourier power spectrum is then used to compute 24 filter bank coefficients, using triangular filters placed on a non-uniform frequency scale, similar to the Bark/Mel scale. We finally take the base 10 logarithm of each filter output and multiply the result by 10, to form a 24-dimensional vector of filter bank coefficients in dB.

4.4. TFPC Filtering

Once spectral vectors have been extracted from a training utterance, we calculate the TFPC corresponding to that sequence using $q = 1, 2, 3$, and keep each time all the components. We then filter the spectral vectors by these components to obtain a new set of feature vectors. Thus, for each speaker, we have a set of components for the filtering, and a sequence of filtered vectors by these components. For comparison, we also compute the first 12 cepstral vectors (c_1 to c_{12}) [8], and the corresponding Δ parameters [2] using 3 or 5 frames. Since the TFPC approach can be applied to any kind of vectors, we also apply it to the sequence of cepstral vectors instead of the sequence of spectral vectors.

4.5. Modeling

Each training sequence is then modeled by a Gaussian mixture model [6, 9] using 8 components and diagonal covariance matrices.

4.6. Test Phase

During the test phase, before calculating the log-likelihood of a test sequence of spectral vectors given a training model, we first filter this sequence by the corresponding time-frequency principal components, which can be interpreted as a projection on a particular sub-space, and we then calculate the log-likelihood. The model which gives the highest log-likelihood will determine the identity of the test utterance.

4.7. Results and Discussion

Table 1 presents the results of our experiments.

Coeff.	static	static + delta		TFPC filtering		
Context	1 vect.	3 vect.	5 vect.	1 vect.	3 vect.	5 vect.
Spec.	21.96	-	-	9.82	**9.11**	9.82
Ceps.	15.71	12.32	**11.43**	10.00	9.29	10.71

Table 1: *Percentage of identification error.*

The reference identification error rate is the one obtained when using cepstral coefficients augmented by the Δ parameters. In our experiments, the best score is obtained with a Δ calculation over 5 vectors (11.43 %).

In both cases (when starting with spectral coefficients or cepstral coefficients), the error rates obtained after application of the TFPC filtering are better, and are lower when the TFPC filtering is applied using 3 vectors.

Finally, the best error rate is obtained when applying the TFPC filtering to spectral coefficients (9.11 %) and outperforms the reference error rate of 20.3 %.

5. CONCLUSION

We have presented a new formalism, called vector filtering of spectral trajectories, which allows to integrate several speech parameterizations under a common formalism. We have also studied a new speech parameterization called time-frequency principal components. Applied in the framework of closed-set text-independent speaker identification, this new approach shows an improvement in the identification error rate of roughly 20 % when compared to the use of the classical cepstral coefficients augmented by their Δ coefficients.

6. FUTURE DIRECTIONS

In our experiments, we have kept all the TFPC extracted. It may be interesting to study the degradation of the performance if we keep only the first components. We may have a more compact representation with a small loss in performance, or even a gain in performance since the last components correspond mainly to noise.

The TFPC filtering can also be tested in the framework of text-independent speaker verification. In that case, it must also be applied to the sequence of vectors used to train the background models. Then, during the test phase, before calculating the likelihood ratio, the sequence of test vectors will be projected on two different sub-spaces, the sub-space corresponding to the claimed identity model, and the sub-space corresponding to the background model.

Finally, the TFPC filtering can more generally be applied to any pattern recognition problem where the task is to discriminate between classes, as in language recognition or speech recognition for instance.

7. REFERENCES

[1] *The European COST 250 Action entitled: Speaker Recognition in Telephony*, 1998. (http://circhp.epfl.ch/polycost).

[2] Sadaoki Furui. Comparison of speaker recognition methods using static features and dynamic features. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):342–350, June 1981.

[3] Yves Grenier. Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur. In *XIèmes Journées d'Etude sur la Parole*, pages 163–171, May 1980. Strasbourg, France.

[4] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[5] Ivan Magrin-Chagnolleau. *Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte*. PhD thesis, École Nationale Supérieure des Télécommunications, January 1997.

[6] Geoffrey J. McLachlan and Kaye E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988.

[7] Ben Milner. Inclusion of temporal information into features for speech recognition. In *Proceedings of ICSLP 96*, 1996.

[8] Alan V. Oppenheim and Ronald W. Schafer. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, 16(2):221–226, June 1968.

[9] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, January 1995.