LANGUAGE RECOGNITION USING TIME-FREQUENCY PRINCIPAL COMPONENT ANALYSIS AND ACOUSTIC MODELING

Michel DUTAT [(1)(2)], Ivan MAGRIN-CHAGNOLLEAU [(3)], Frédéric BIMBOT [(3)]

[(1)] LSCP / CNRS URA 8554, 54 Boulevard Raspail, 75270 Paris cedex, France – European Union.
[(2)] ENST / CNRS URA 820, Dépt. Signal et Image, 46 Rue Barrault, 75634 Paris cedex 13, France – European Union.
[(3)] IRISA (CNRS & INRIA), Campus universitaire de Beaulieu, 35042 Rennes cedex, France – European Union.
dutat@lscp.ehess.fr - ivan@ieee.org - bimbot@irisa.fr

ABSTRACT

Time-Frequency Principal Component (TFPC) is a speech parameterization technique based on a principal component analysis applied to acoustic feature parameters augmented by their time context. In this paper, we investigate on the performance of TFPC in the framework of automatic language recognition. In our experiments, identification rate is improved compared to the use of the conventional cepstral coefficients augmented by their Δ coefficients.

1. INTRODUCTION

Several speech parameterizations have been tested in the past for language recognition. However, best results are usually achieved by classical acoustic features, namely cepstral coefficients plus dynamic information using approximations of their first derivative (Δ coefficients).

In this work, we study the benefit of using TFPC (Time-Frequency Principal Components) [4, 5] for incorporating dynamic information in a data-driven manner. The approach can be viewed as a language-dependent selection of acoustical information both in the time and frequency dimensions. This selection is done using time-frequency vector filters, operating on the sequence of acoustic features, the coefficients of which are estimated on a training corpus. The underlying hypothesis is that a filtered utterance is better modeled when the filter is the one corresponding to its language.

After presenting in more details the modeling and recognition schemes, we apply them to a language identification task. In our experiments, we observe an improvement of the identification rate compared to the one obtained with a reference experiment, using a cepstral parameterization and the Δ coefficients.

2. PRESENTATION OF THE METHOD

2.1. Classical Approach

Acoustic modeling of a language consists in analysing, capturing, and modeling the sound structure of this language. In statistical approaches, this is done by using a great number of training utterances covering a large variety of speakers. In a first step, the utterances have to be parameterized into acoustic features corresponding to successive speech frames. Then, the speech frame distribution is modeled by a statistical model, usually a Hidden Markov Model (HMM).

Figure 1 presents the classical procedure of a training phase. The set of training utterances $E_{app}^{(i)}$ of the language i is transformed through an acoustical analysis into a sequence of parameter vectors $\{\mathbf{x}_t\}^{(i)}$. In training mode, this sequence is used for the estimation of the parameters $\kappa^{(i)}$ of the model of language i. Once the

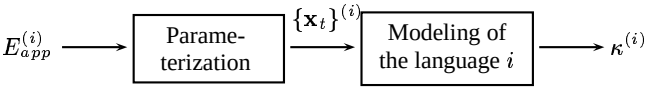

Figure 1: Language recognition : classical training procedure.

parameters of all models have been calculated, n statistical models are obtained, each of them representing one language. The recognition phase then consists in parameterizing the test utterance (as was done in the training phase) and calculating a score with each of the n models. In the case of language identification, a decision algorithm combines the n scores in order to select the most probable language.

2.2. TFPC Approach

The method that we present here was originally proposed in [4, 5]. Its goal is to use a language-dependent parametrization of the utterance, with the goal to make it thus more easy to recognize. Instead of using a common acoustic feature analysis, the sequence of cepstral coefficients of the speech utterance is filtered by a time-frequency filter depending on the language.

The training phase is therefore composed of two steps. The first one consists in a principal component analysis of the training data, where the principal component analysis is applied to sequences of consecutive vectors, so that not only the static information is taken into consideration but also the dynamic information. A language-dependent projection matrix is thus obtained (also called filtering matrix). The second step is the estimation of the statistical model $\lambda^{(i)}$ of the language i using the vector-filtered training acoustic features, with the filter corresponding to the language of the training utterance.

Figure 2 shows the training phase for the language i. The set of training utterances $E_{app}^{(i)}$ is transformed by an acoustic analysis

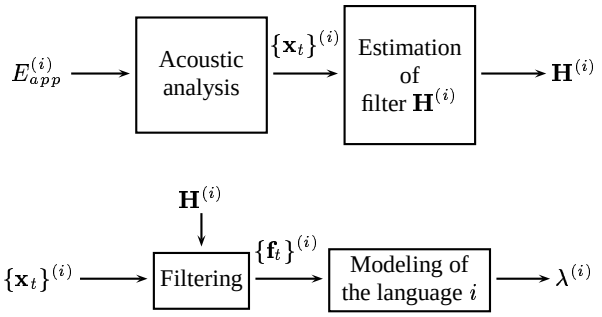

Figure 2: Language recognition : training procedure with TFPC.

(spectral, cepstral, linear prediction, ...) into a sequence $\{\mathbf{x}_t\}^{(i)}$ of p-dimensional vectors. This sequence is then used to compute the filtering matrix $\mathbf{H}^{(i)}$. Once the filtering matrix has been calculated, the sequence $\{\mathbf{x}_t\}^{(i)}$ is filtered by this matrix and it provides a new sequence of r-dimensional vectors $\{\mathbf{f}_t\}^{(i)}$ used to estimate the model $\lambda^{(i)}$ for the language i.

2.3. Construction of the Filter

For a given language and corresponding training data, the TFPC approach extracts the representation which maximizes the inertia of the data in a language-characteristic sub-space. The underlying idea is to enhance static and dynamic language-dependent correlations which may exist between the acoustic features and which may characterize the language. In order to capture the dynamic information, we take the time context of each vector into consideration.

The successive steps for the construction of the filter are the following : let $\{\mathbf{x}_t\}$ be the sequence of p-dimensional acoustic vectors extracted from the training utterances. From this sequence, we extract the expanded vectors $\mathbf{X}_{t-q}^{t+q}$:

$$\mathbf{X}_{t-q}^{t+q} = \begin{bmatrix} \mathbf{x}_{t+q}^* \\ \vdots \\ \mathbf{x}_t^* \\ \vdots \\ \mathbf{x}_{t-q}^* \end{bmatrix} \qquad \text{with } \mathbf{x}_t^* = (\mathbf{x}_t - \overline{\mathbf{x}})$$

Then we compute the covariance matrix of the expanded vectors (contextual covariance matrix). This matrix can be calculated by computing the lagged covariance matrices R_q :

$$\mathsf{R}_q = \frac{1}{T}\sum_{t=q+1}^{T}(\mathbf{x}_t - \overline{\mathbf{x}})(\mathbf{x}_{t-q} - \overline{\mathbf{x}})^\mathsf{T} \quad \text{with} \quad \overline{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$$

The number of lagged covariance matrices corresponds to the order of the TFPC analysis. The contextual covariance matrix is then obtained by combining the lagged covariance matrices in a block-Toeplitz matrix $\mathbf{R}_{2q+1}$.

$$\mathbf{R}_{2q+1} = \begin{bmatrix} \mathsf{R}_0 & \mathsf{R}_1 & \cdots & \mathsf{R}_{2q} \\ \mathsf{R}_1^\mathsf{T} & \mathsf{R}_0 & \cdots & \mathsf{R}_{2q-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{R}_{2q}^\mathsf{T} & \mathsf{R}_{2q-1}^\mathsf{T} & \cdots & \mathsf{R}_0 \end{bmatrix}$$

The principal component analysis of the contextual covariance matrix is obtained by calculating its eigenvalues and eigenvectors, and ordering the eigenvalues (and the corresponding eigenvectors) in decreasing order. The sub-space generated by the eigenvectors corresponding to the highest eigenvalues has the highest inertia. By construction, all the directions of the eigenvectors are orthogonal. The principal components associated with the smallest eigenvalues often correspond to noise. Therefore, they are usually removed. Other strategies can be adopted to select the components [3].

The filtering matrix is finally obtained by transposing the matrix containing the selected eigenvectors. For instance, if the 5 first components and the components 10 to 12 are to be kept, the filtering matrix $\mathbf{H}$ will be :

$$\mathbf{H} = [\mathbf{v}_1 \cdots \mathbf{v}_5\ \mathbf{v}_{10} \cdots \mathbf{v}_{12}]^\mathsf{T}$$

The filtering of an utterance is ultimately obtained by calculating the convolution between the matrix $\mathbf{H}$ and the expanded vectors extracted from the utterance. For instance, when only the first r principal components are kept, we have :

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

And the filtered vectors are obtained as :

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

2.4. Recognition Phase

Once the training phase is completed, each language is characterized by a filtering matrix and a statistical model. Identifying the language of a test utterance is carried out by filtering the utterance using the filtering matrix of each language, and then computing for each corresponding model a likelihood measure.

This procedure is illustrated on Figure 3. A first acoustic analysis transforms the test utterance E_{test} into a sequence of vectors $\{\mathbf{x}_t\}$. For each language, this sequence is filtered. Then an estimation of the conditional probability $P\left(\{\mathbf{f}_t\}^{(i)}|\lambda^{(i)}\right)$ of observing $\{\mathbf{x}_t\}$ knowing its language is $L^{(i)}$ is calculated. The language which gives the maximum likelihood is finally chosen.

With this approach, several parameters can vary:

- the initial acoustic analysis of the utterances (filterbank analysis, LPCC, cepstral coefficients, ...);
- the order of the TFPC analysis, that is, the number of frames taken into account;
- the choice of the principal components to keep for the construction of the filtering matrix.

In the next sections, we present some results obtained by varying some of these parameters. We also present results obtained with a reference parameterization, the cepstral parameterization, in order to compare the TFPC approach to a more conventional one.

3. EXPERIMENTS

3.1. Objectives

Our objective is to evaluate the quality of the new parameterization on a language identification task. We test two language recognition

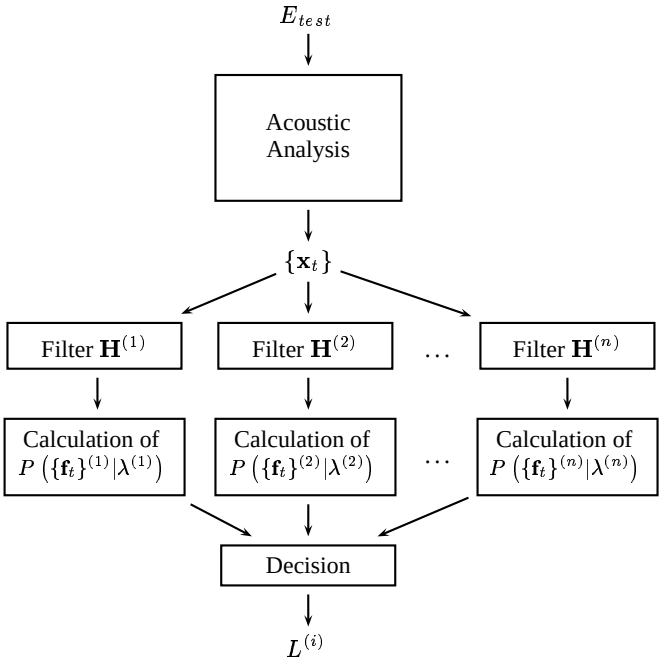

Figure 3: Language recognition : test procedure using the TFPC approach.

approaches which only differ by the parameterization step, one using the classical cepstral parameterization (reference experiment), the other using the TFPC parameterization.

3.2. Corpus

The database used for our experiments is a subset of the OGI MLT-S (Oregon Graduate Institute Multi-language Telephone Speech) database [6]. We selected four languages: English, French, Japanese, and Spanish. For each language, we use 80 utterances of 15 seconds in average pronounced by 20 speakers for the training phase. This corresponds approximately to 20 min. of speech for each language. For the test, we use, for each language, 54 utterances of 15 seconds in average pronounced by 12 (other) speakers. We use therefore a total of 216 test utterances.

3.3. Characteristics of the Reference Experiment

The acoustic model consists in an Ergodic Hidden Markov Model (E-HMM) for each language. The characteristics of this model are as follows :

Structure	Ergodic
State number	24
Number of Gaussians by state	2
Covariance matrices	Diagonal

For the parameterization of the speech signal, we use the cepstral coefficients rather than the spectral coefficients according to a preliminary study [1]. The characteristics of this analysis are the following :

Type of parameterization	MFCC [1]
Frequency scale	Mel scale
Number of coefficients	12 (optionally + 12 Δs)
Size of the analysis window	30 ms
Shift between two windows	10 ms
Type of window	Hamming
Cepstral mean subtraction	Yes

Some experiments with TFPCs take into account the temporal context. In order to make the experiments comparable, we did similarly for the reference experiments by using, in this case, the Δ parameters, calculated as follows [2] :

$$\Delta c_t^i = \frac{\sum_{k=1}^{q} k(c_{t+k}^i - c_{t-k}^i)}{2 \cdot \sum_{k=1}^{q} k^2} \qquad \text{avec} \quad i = 1, 2 \cdots, p$$

where q is half the size of the temporal context.

3.4. Experiments with TFPC

The acoustic model used for these experiments is exactly the same as for the reference experiment (Ergodic HMM). We also use, as initial acoustic analysis, the cepstral parameterization as described above.

Then, we apply the TFPC analysis to the cepstral vectors for various sizes of context and, in each case, we vary the number of principal components which are kept. We present in the next sections the most significant results for the order 0, which corresponds to no context, and for the order $q = 4$, which provides the best results. The order 4 corresponds to a time context of 9 frames (110 ms).

3.4.1. Order 0

The TFPC analysis at order $q = 0$ does not use any time context. The contextual covariance matrix is therefore the usual covariance matrix R_0, with dimension 24×24. The filtering matrix has the same dimension or is smaller, and the filtered vectors are 24-dimensional or smaller. In that case, the reference experiment uses 24-dimensional MFCCs and no Δ parameters. It yields an identification rate of **53.1 %**.

The table below shows the average identification score in percentage for various numbers of TFPC components. The last row shows the relative gain compared to the reference experiment. It can be noticed that the system with the TFPC analysis performs better than the reference experiment, and that 16 is the optimal number of components to keep in this case. The TFPC approach seems therefore to capture language-dependent information in the first principal components, whereas the last components seem to be less meaningful for the task.

TFPC components kept	12	16	20	24
Identification rate (in %)	54.3	**58.0**	55.8	53.5
Relative gain to ref. (in %)	2.3	**9.3**	5.1	0.7

3.4.2. Order 4

The order $q = 4$ corresponds to a context of 9 frames (110 ms). The dimension of the contextual covariance matrix $\mathbf{R}_9$ is now

216×216. The reference experiment includes 12 MFCC plus 12 Δ MFCC calculated with $q = 4$. The corresponding identification rate is **60.3 %**. The table below shows the results for the TFPC experiment. Here too, a noticeable improvement is observed as compared to the reference experiment, and the optimal number of components in that case is 24.

TFPC components kept	20	24	28	32
Identification rate (in %)	58.6	**63.5**	53.6	60.8
Relative gain to ref. (in %)	-3.0	**5.2**	-11.2	0.7

3.5. Results using a Voice Activity Detector

We carried out the same experiments after implementing a voice activity detector as a pre-processing. The detector is based on a bi-Gaussian modeling of the energy, and on a dynamic threshold calculation. A post-processing is added to penalize the detection in silence or noise areas and to reinforce the detection in the speech areas. The results are presented in the next sections.

3.5.1. Order 0

For the order $q = 0$, the identification rate for the reference experiment increases to **59.1 %**. For the TFPC experiments, we obtain:

TFPC components kept	16	20	24
Identification rate (in %)	59.1	62.6	**63.6**
Relative gain to ref. (in %)	0.0	6.0	**7.7**

As could be expected, the results are globally better when we use a voice activity detector. It is also interesting to note that, in that case, all the coefficients seem to contain useful information, although the first 16 coefficients seem to contain already a great part of this information. Finally, we note that the system with the TFPC analysis performs again better than the reference system.

3.5.2. Order 4

With a time context of 9 frames, the reference experiment reaches an identification rate of **69.2 %**. The results of the TFPC experiments are given in the following table:

Composantes	24	36	44	50
Identification rate (in %)	65.7	69.7	**75.3**	74.8
Relative gain to ref. (in %)	-5.1	0.7	**8.8**	8.0

This time, we need to take more components (namely 44) to achieve the best performance. If we take only 24 components as previously, the score is significantly worse than the reference (using also 24 coefficients). The system using the TFPC analysis with 44 components yields an identification score of **75.3 %**, which is better than the reference system, but at the expense of a higher dimensionality. However, an additional experiment using 24 MFCC + 24 Δ-MFCC (i.e 48 coefficients in total) gives an identification rate of **68.7 %** only, which shows that the performance improvement observed with 44 TFPCs is not owed to the higher dimensionality, but to the acoustic features themselves.

4. SUMMARY AND PERSPECTIVES

In this paper, we have presented the use of the Time-Frequency Principal Component (TFPC) analysis, for performing a language-dependent vector-filtering on acoustic feature parameters. The Principal Component Analysis is applied to the contextual covariance matrix which is the covariance matrix of a sequence of vectors augmented by their time context.

The filters can be interpreted as projection matrices on language-characteristic sub-spaces. When we apply the TFPC parameterization to language identification, the best configuration of our system gives an identification score of **75.3 %**, outperforming by approximately 9 % a reference system using 12 cepstral and 12 Δ-cepstral coefficients.

We now intend to study the influence of such a parameterization on the statistics of the state *sequences* in the Hidden Markov Models. This is particularly important in the perspective of identifying languages via typical regularities of acoustic events.

5. REFERENCES

[1] Michel Dutat. Reconnaissance automatique de la langue parlée. Rapport d'avancement de thèse, Ecole Nationale Supérieure des Télécommunications, 1997.

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

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

[5] Ivan Magrin-Chagnolleau, Geoffrey Durou, and Frédéric Bimbot. Application of time-frequency principal component analysis to text-independent speaker identification. Submitted to IEEE Transactions on Speech and Audio Processing.

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. Technical report, Center for Spoken Language Understanding Oregon Graduate Institute of Science and Technology., Portland, 1993.