INDEXING TELEPHONE CONVERSATIONS BY SPEAKERS USING TIME-FREQUENCY PRINCIPAL COMPONENT ANALYSIS

Ivan Magrin-Chagnolleau and Frédéric Bimbot

IRISA (CNRS & INRIA) - Campus universitaire de Beaulieu - 35042 Rennes cedex - France - European Union
ivan@ieee.org - bimbot@irisa.fr

ABSTRACT

In this paper, we present an algorithm for the tracking of target speakers in telephone conversations. Speaker tracking consists in retrieving, in an audio recording, segments which have been uttered by a target speaker. We also compare two speech analysis techniques. The first one is the time-frequency principal component analysis. It is a new speech analysis technique based on the extraction of the principal components of the contextual covariance matrix, which is the covariance matrix of feature vectors expanded by their time context. The other one is the classical cepstral analysis. Experiments are carried out on a subset of the Switchboard database.

1. INTRODUCTION

Speaker tracking is an emerging task in multimedia data processing. It consists in tracking one or several target speakers in an audio recording. This new task has become more and more important with the increase of multimedia data available through Internet. Speaker tracking is one of the numerous new tools used to segment, classify, and organize these data, and therefore to access them faster and more effectively. This emerging problem has been reported on only recently [13, 9, 1, 2, 5, 14].

In this paper, we present a tracking algorithm based on a log-likelihood ratio calculation and a multiple thresholds segmentation algorithm [13, 9]. We also compare two speech analysis techniques in the framework of speaker tracking. The first one is a new speech analysis technique called *time-frequency principal component (TFPC)* analysis [6, 7, 8]. TFPC analysis consists in calculating the principal components of the contextual covariance matrix, which is the covariance matrix of a sequence of vectors (for instance spectral vectors) expanded by their time context. We compare this new analysis to a classical cepstral analysis, using cepstral coefficients [11] augmented by their Δ parameters [3].

The experiments reported in this paper are carried out on a subset of the Switchboard database. This subset is a part of the development data of the NIST 2000 speaker recognition evaluation.

2. TIME-FREQUENCY PRINCIPAL COMPONENTS

TFPC analysis consists in extracting time-frequency patterns which are characteristic of a whole sequence of training vectors, that is, to summarize the evolution of a sequence of training vectors by a few short sequences extracted from the entire sequence. The original sequence has therefore 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 original 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. In this paper, the original vectors are spectral vectors. But the same procedure can be used on any kind of vectors. The only requirement is that the vectors are indexed by time.

Let $\{\mathbf{x}_t\}_{1 \leq t \leq M}$ denote a sequence of 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:

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

This matrix is block-Toeplitz, and its dimension is $(2q+1)\,p \times (2q+1)\,p$.

Let define the sequence of vectors $\mathbf{x}_t$ between time $t-q$ and $t+q$ by:

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

This work has been done in the framework of the first phase of the AGIR project, funded by the French RNRT Programme.

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

The matrix $\mathbf{X}_{2q+1}$ can be interpreted as the covariance matrix of the vectors $\{\mathbf{X}_{t-q}^{t+q}\}_{1 \leq t \leq M}$, and can therefore be called *contextual covariance matrix*.

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:

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

with:

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

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

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

Once the TFPC have been calculated, we choose some of them to build a filtering matrix. For instance, if we choose to keep the components $\{\mathbf{v}_1, ..., \mathbf{v}_6, \mathbf{v}_{14}, ..., \mathbf{v}_{22}\}$, the filtering matrix for the corresponding class will be:

$$\mathbf{H} = \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_6^T \\ \mathbf{v}_{14}^T \\ \vdots \\ \mathbf{v}_{22}^T \end{bmatrix}$$

3. SEGMENTATION ALGORITHM

Given a telephone conversation, speaker tracking consists in detecting the segments in which a target speaker is speaking, that is, to estimate the beginning and the end of these segments. First, spectral vectors are extracted from the telephone conversation, using the spectral analysis described in the next section.

Let $\{\mathbf{x}_t\}_{1 \leq t \leq T}$ denote the sequence of spectral vectors.

TFPC Analysis. In the case of TFPC analysis, spectral vectors are then filtered by the TFPC extracted from the training data of the target speaker in order to obtain a new sequence of vectors, $\{\mathbf{x}_t^T\}$. The spectral vectors are also filtered by the TFPC extracted from the world data, and we obtain another sequence of filtered vectors, $\{\mathbf{x}_t^W\}$.

The world data contain speech from a population of speakers from the same gender as the target speaker, and using the same type of handset[1]. This population is supposed to be representative of the world population in term of speaker variability. The model extracted from the world data is used to normalize the likelihood score during the tracking phase.

Let λ_{tfpc}^T and λ_{tfpc}^W denote respectively the model for the target speaker and the world model in the case of TFPC analysis. We then calculate the likelihood of $\mathbf{x}_t^T$ given the model λ_{tfpc}^T, denoted by $\mathcal{L}(\mathbf{x}_t^T|\lambda_{tfpc}^T)$, the likelihood of $\mathbf{x}_t^W$ given the model λ_{tfpc}^W, denoted by $\mathcal{L}(\mathbf{x}_t^W|\lambda_{tfpc}^W)$, and finally the likelihood ratio, denoted by $\mathcal{R}(\mathbf{x}_t^T; \mathbf{x}_t^W|\lambda_{tfpc}^T; \lambda_{tfpc}^W)$, whose logarithm is expressed by:

$$\log \mathcal{R}(\mathbf{x}_t^T; \mathbf{x}_t^W|\lambda_{tfpc}^T; \lambda_{tfpc}^W) = \log \mathcal{L}(\mathbf{x}_t^T|\lambda_{tfpc}^T) - \log \mathcal{L}(\mathbf{x}_t^W|\lambda_{tfpc}^W)$$

Cepstral Analysis. In the case of cepstral analysis, spectral vectors are transformed through cosine functions in order to form cepstral vectors [11]. Then, these cepstral vectors are augmented by their Δ parameters [3]. We finally obtain the sequence $\{\mathbf{c}_t\}$ containing cepstral coefficients and Δ-cepstral coefficients. Let λ_{ceps}^T and λ_{ceps}^W denote respectively the model for the target speaker and the world model in the case of cepstral analysis. We then calculate the likelihood of $\mathbf{c}_t$ given the model λ_{ceps}^T, denoted by $\mathcal{L}(\mathbf{c}_t|\lambda_{ceps}^T)$, the likelihood of $\mathbf{c}_t$ given the model λ_{ceps}^W, denoted by $\mathcal{L}(\mathbf{c}_t|\lambda_{ceps}^W)$, and finally the likelihood ratio, denoted by $\mathcal{R}(\mathbf{c}_t|\lambda_{ceps}^T; \lambda_{ceps}^W)$, whose logarithm is expressed by:

$$\log \mathcal{R}(\mathbf{c}_t|\lambda_{ceps}^T; \lambda_{ceps}^W) = \log \mathcal{L}(\mathbf{c}_t|\lambda_{ceps}^T) - \log \mathcal{L}(\mathbf{c}_t|\lambda_{ceps}^W)$$

Smoothing. Before applying the segmentation algorithm, smoothing is needed to attenuate variations of the log-likelihood ratio. The smoothing is an arithmetic mean of a specified number of consecutive values of the log-likelihood ratio. Two parameters define the smoothing: the number of values used for the average calculation, denoted by τ; and the delay between two calculations, that is, the number of feature vectors between two average calculations, denoted by d. In our experiments, τ was set to 30 vectors (320 ms) and d to 5 vectors (70 ms).

Segmentation Algorithm. A segmentation algorithm using multiple thresholds [13, 9] is then applied on the average values previously calculated. The algorithm is illustrated on Figure 1.

On this figure, the smoothed log-likelihood ratio is plotted in solid line. When its value becomes higher than the threshold θ_0, the current time is recorded as the beginning of a possible segment. But the beginning of the segment is marked only if the value of the smoothed log-likelihood ratio becomes also higher than the threshold θ_1. This procedure allows us to avoid the detection of a segment when the score fluctuates just a little bit around θ_0, and therefore reduces the number of false alarms. The same principle is applied for the detection of the end of a segment, with the two thresholds θ_2 and θ_3. Figure 1 shows an example with the values

[1] Gender and handset are automatically detected during the tracking phase.

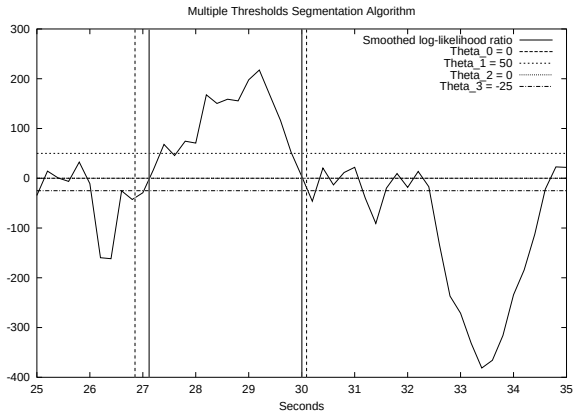

Figure 1: Illustration of the segmentation algorithm using multiple thresholds. The vertical solid lines are the boundaries detected by the segmentation algorithm. The vertical dash lines are the real boundaries of the segment.

$\theta_0 = \theta_2 = 0$, $\theta_1 = 50$, and $\theta_3 = -25$.

Finally, the minimum duration for an estimated segment is set to 2.5 seconds (every segment whose duration is smaller is omitted), and the minimum interval between two consecutive segments to 1 second (two consecutive segments are merged if the interval between them is smaller). The segmentation algorithm provides the estimated beginning and end times for the target speaker segments.

4. EXPERIMENTS

Task. Given a telephone conversation between two persons, and a target speaker, the speaker tracking task consists in tracking the target speaker in the telephone conversation. The target speaker may or may not be one of the two speakers engaged in the conversation.

Database. The database used for these experiments was a subset of the Switchboard 2 - Phase 3 corpus collected by the Linguistic Data Consortium (LDC). The whole corpus consists of 5 minute telephone conversations involving 640 speakers, most of them being college students in the southern United States. The population of speakers used for the experiments was composed of 120 speakers (80 females and 40 males). For each of them, two sentences of approximately 1 minute of speech were available for the training phase. The test data was composed of 66 telephone conversations, which were tested against 4 hypothetic target speakers, 2 of them being the 2 speakers of the conversation.

Data from the Switchboard database was also used to build world models. These models were used to normalize the log-likelihood scores during the tracking phase. These models were gender-dependent and handset-dependent. Therefore, we had one world model for females using electret handset, one world model for females using carbon handset, and so on. Data for these models were obtained from a subset composed of 100 speakers (25 females using electret handset, 25 females using carbon handset, 25 males using electret handset, 25 males using carbon handset).

Spectral Analysis. For each speaker of the training dataset, the two training sentences were concatenated. Then, the speech signal was decomposed in frames of 20 ms at a frame rate of 10 ms. A Hamming window was applied to each frame. The signal was pre-emphasized with a coefficient 0.95. For each frame, a fast Fourier transform was computed and provided 256 square modulus values representing the short term power spectrum in the 0-4 kHz band. This Fourier power spectrum was then used to compute 24 filterbank coefficients, using triangular filters placed on a linear frequency scale in the bandwidth 300-3400 Hz. We finally took the base 10 logarithm of each filter output and multiplied the result by 10, to form a 24-dimensional vector of filterbank coefficients in dB. The same spectral analysis was applied to the data of the world model.

TFPC and Cepstral Analysis. Once spectral vectors have been extracted from a training utterance, we calculated the TFPC corresponding to that sequence using the value $q = 2$ (which corresponds to a context of 5 spectral vectors), and kept only the 24 first components. We then filtered the spectral vectors by these components to obtain a new set of feature vectors. Therefore, for each speaker, we had a set of components for the application of the TFPC filtering, and a sequence of vectors filtered by these components. The TFPC analysis was also applied to the spectral vectors extracted from the world data.

We also extracted 12 cepstral coefficients [11] from the spectral vectors, and computed their Δ parameters [3], obtaining 24-dimensional vectors.

Modeling. Each training sequence was then modeled by a Gaussian mixture model (GMM) [15, 12] using 256 or 512 components and diagonal covariance matrices. The Gaussian mixture models were calculated using an expectation maximization (EM) algorithm, initialized by a vector quantization (VQ) algorithm. The same modeling was used for each world model using 512 components and diagonal covariance matrices.

Tracking Phase. Each test consisted in a telephone conversation between two unknown speakers and the identity of one target speaker that we wanted to track in the telephone conversation. We applied the segmentation algorithm described in Section 3.

5. RESULTS AND DISCUSSION

The performance of the system was measured by a DET curve [10]. This representation is a way to show all of the possible operating points of a system (false alarm rate vs. miss rate) in a scale which makes the result curves rather linear. For this task, the false alarm rate (or false acceptance rate) and the miss rate (or false rejection rate) were defined in the following way:

$$\mathcal{R}_{FA} = \frac{\text{Number of non-target frames labeled as target}}{\text{Number of non-target frames}}$$

$$\mathcal{R}_{MI} = \frac{\text{Number of target frames labeled as non-target}}{\text{Number of target frames}}$$

Results are reported on Figure 2 and show the performance of the system for the two speech parameterizations and for two model sizes, 256 or 512 components.

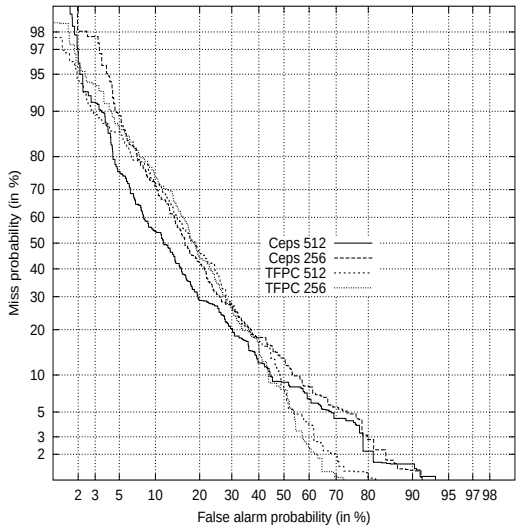

Figure 2: Comparison between a system based on the TFPC analysis and a system based on the cepstral analysis for two model sizes (256 or 512 components).

The first remark is that the four configurations of the system perform very similarly. The cepstral-based systems are better in the middle range of the DET curve, in particular around the EER (Equal Error Rate) which is approximately 26 % for the system using 512 components. But the TFPC-based systems perform slightly better in the extremities of the DET curve, particularly when the false alarm probability is high and the miss probability is low. It is also interesting to notice that, in the case of the TFPC-based systems, the two model sizes give similar performances.

6. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have used a speaker tracking system based on a log-likelihood ratio calculation and a multiple thresholds segmentation algorithm to track target speakers in telephone conversations. Two speech analysis techniques have been tested. The first one, the time-frequency principal component analysis, is a new speech analysis technique based on the extraction of the principal components of the contextual covariance matrix, which is the covariance matrix of feature vectors expanded by their time context. The second one is the classical cepstral analysis.

The results reported show very similar performance between the two speech parameterizations. The cepstral-based system performs slightly better in the middle range of the DET curves, which corresponds to the area where the two types of error rates are close to each other. The TFPC-based system performs better on the extremities of the DET curves, which corresponds to the areas where one error rate is much higher than the other one. If the chosen operating point is in the middle part of the DET curve, one will choose a cepstral-based system. On the contrary, if the chosen operating point is in the extremities of the DET curve, one will rather choose a TFPC-based system.

The performances shown in this paper correspond to a raw calculation of the log-likelihood ratios. It may be interesting to use a normalization technique like the z-norm or the h-norm in order to improve the performances. These normalization techniques are based on a modeling of the score distributions. It has been shown to work pretty well for speaker verification [1]. The authors intend to test these normalization techniques in the framework of speaker tracking.

An other perspective to this work is the use of a Maximum A Posteriori (MAP) training algorithm, as it is presented in [14], instead of a Maximum Likelihood (ML) training algorithm.

7. REFERENCES

[1] Robert B. Dunn, Douglas A. Reynolds, and Thomas F. Quatieri. Approaches to speaker detection and tracking in multi-speaker audio. *Digital Signal Processing*, 10(1-3), April 2000.

[2] Corinne Fredouille, Jean-François Bonastre, and Teva Merlin. AMIRAL: A block-segmental multi-recognizer architecture for automatic speaker recognition. *Digital Signal Processing*, 10(1-3), April 2000.

[3] 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.

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

[5] Johan Koolwaaij and Lou Boves. Local normalization and delayed decision making in speaker detection and tracking. *Digital Signal Processing*, 10(1-3), April 2000.

[6] Ivan Magrin-Chagnolleau and Geoffrey Durou. Time-frequency principal components of speech: application to speaker identification. In *Proceedings of EUROSPEECH 99*, pages 759–762, September 1999. Budapest, Hungary.

[7] Ivan Magrin-Chagnolleau and Geoffrey Durou. Application of time-frequency principal component analysis to speaker verification. *Digital Signal Processing*, 10(1-3), April 2000.

[8] 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*.

[9] Ivan Magrin-Chagnolleau, Aaron E. Rosenberg, and S. Parthasarathy. Detection of target speakers in audio databases. In *Proceedings of ICASSP 99*, pages 821–824, March 1999. Phoenix, Arizona, United States.

[10] A. Martin et al. The DET curve in assessment of detection task performance. In *Proceedings of EUROSPEECH 97*, volume 4, pages 1895–1898, September 1997. Rhodes, Greece.

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

[12] 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.

[13] Aaron E. Rosenberg, Ivan Magrin-Chagnolleau, S. Parthasarathy, and Qian Huang. Speaker detection in broadcast speech databases. In *Proceedings of ICSLP 98*, December 1998.

[14] Mouhamadou Seck, Raphaël Blouet, and Frédéric Bimbot. The IRISA / ELISA speaker detection and tracking systems for the NIST'99 evaluation campaign. *Digital Signal Processing*, 10(1-3), April 2000.

[15] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and sons, 1985.