# A SPEAKER TRACKING SYSTEM BASED ON SPEAKER TURN DETECTION FOR NIST EVALUATION

*Jean-François Bonastre*[1], *Perrine Delacourt*[2*], *Corinne Fredouille*[1],
*Teva Merlin*[1†], *Christian Wellekens*[2]

[1] LIA/CERI Université d'Avignon, France
{jean-francois.bonastre,corinne.fredouille,teva.merlin}@lia.univ-avignon.fr
[2] Institut Eurécom, Sophia Antipolis, France
{perrine.delacourt,christian.wellekens}@eurecom.fr

## ABSTRACT

A speaker tracking system (STS) is built by using successively a speaker change detector and a speaker verification system. The aim of the STS is to find in a conversation between several persons (some of them having already enrolled and other being totally unknown) target speakers chosen in a set of enrolled users. In a first step, speech is segmented into homogeneous segments containing only one speaker, without any use of a priori knowledge about speakers. Then, the resulting segments are checked to belong to one of the target speakers. The system has been used in a NIST evaluation test with satisfactory results.

## 1. INTRODUCTION

Speaker tracking is an important issue in multimedia applications requiring analysis of spoken documents. An important problem is to detect where a given speaker is intervening in a discussion. In this paper, we propose a Speaker Tracking System (STS) which consists in two main steps. First, a speaker segmentation process detects all speaker changes. This process works with no use of a priori knowledge about the sought speakers. Then, after segmentation, a classical speaker verification system is applied on each segment to determine whether it has been uttered by the target speaker, for which a model is available from previous enrollments. This approach can be seen as a speaker spotting technique since some speakers in the analyzed documents could be unknown from the speaker verification because they had never showed up in previous recorded sessions: for instance a well known politician can be tracked in a discussion with several unidentified persons. Politician's model has been created earlier from available discourses while the laymen have never enrolled.

A simple and efficient Speaker Tracking System is built from better and better mastered techniques as speaker segmentation and speaker verification. It has been tested during the NIST-1999 evaluation campaign [1]. The test consisted in detecting in a conversation between two speakers, one called target speaker being enrolled, the contributions of the target speaker. In our system, no use was made of the number of different speakers (only 2 in the NIST evaluation).

Figure 1 shows our STS. It is divided into three functional blocks. The first one is a front-end processing and corresponds to the standard ELISA consortium[2] parameterization module ([3]). The second block consists in segmenting the parameterized signal in homogeneous portions containing utterances of a single speaker. This part will be fully detailed in section 2. Then, the resulting segments are used in the verification module which constitutes the third block: verification decision is taken on each segment. This process will be described in section 3. Section 4 presents and comments the experiments. Finally, concluding remarks and possible tracks to improve our system are given in section 5.

## 2. FRONT-END SEGMENTATION

The use of a front-end segmentation before the verification process relies on the assumption that a speaker verification score is more reliable when it is computed

[*] The financial support of this project from the Centre National d'Etudes des Télécommunications (CNET) under the grant n$^o$ 98 1B is gratefully acknowledged here.

[†] This work was supported by the LIARMA Project (LIA/Avignon and RMA/Brussels).

[1] http://www.itl.nist.gov/iaui/894.01/test.htm

[2] The ELISA consortium is composed of European research laboratories working on a shared reference platform for the evaluation of speaker recognition systems. These labs are: ENST (France), EPFL (Switzerland), IDIAP (Switzerland), IRISA (France), LIA (France), RIMO — Rice (USA) and Mons (Belgium) —, RMA (Belgium), VUTBR (Czech Republic).

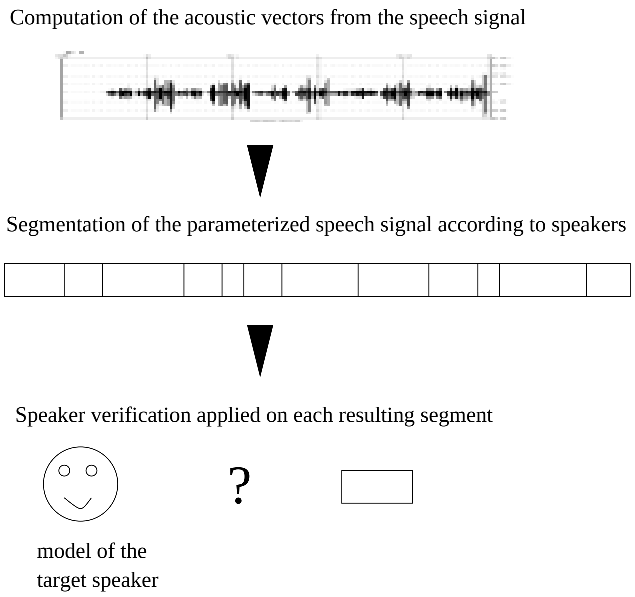
Computation of the acoustic vectors from the speech signal
Segmentation of the parameterized speech signal according to speakers
Speaker verification applied on each resulting segment
?
model of the
target speaker

Figure 1: *Our Speaker Tracking System*

on a large number of frames (acoustic vectors) than on a few frames (see for example [8]). Thus, the goal of the front-end segmentation is to split a parameterized signal into speaker homogeneous segments: the resulting segments should be related to a single speaker and as long as possible. The speaker-based segmentation is performed without speaker models. It relies on the detection of speaker turns in the parameterized signal (the target speaker model is not used in this step).

### 2.1. Detection of one speaker turn

Given two adjacent portions of parameterized signal (sequences of acoustic vectors) $\mathcal{X}_1 = \{x_1, ..., x_i\}$ and $\mathcal{X}_2 = \{x_{i+1}, ..., x_{N_\mathcal{X}}\}$, we consider the following hypothesis test for a speaker turn at time $i$:
• $H_0$: both portions are generated by the same speaker. Then the reunion of both portions is modeled by a multi-Gaussian process $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \sim \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X})$
• $H_1$: each portion is pronounced by a different speaker. Then each portion is modeled by a multi-Gaussian process $\mathcal{X}_1 \sim \mathcal{N}(\mu_{\mathcal{X}_1}, \Sigma_{\mathcal{X}_1})$ and $\mathcal{X}_2 \sim \mathcal{N}(\mu_{\mathcal{X}_2}, \Sigma_{\mathcal{X}_2})$

The Generalized Likelihood Ratio (GLR) between the hypothesis $H_0$ and $H_1$ is defined by:

$$R = \frac{L(\mathcal{X}, \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X})}{L(\mathcal{X}_1, \mathcal{N}(\mu_{\mathcal{X}_1}, \Sigma_{\mathcal{X}_1}).L(\mathcal{X}_2, \mathcal{N}(\mu_{\mathcal{X}_2}, \Sigma_{\mathcal{X}_2})}$$

The GLR has been used in [6, 7] for speaker verification and has proved its efficiency. The distance $d_\mathrm{R}$ is computed from the logarithm of the previous expression: $d_\mathrm{R} = -\log R$.

A high value of $R$ (i.e. a low value of $d_\mathrm{R}$) signifies that the one multi-Gaussian modeling (hypothesis $H_0$) fits best the data. By contrast, a low value of $R$ (i.e. a high value of $d_\mathrm{R}$) indicates that the hypothesis $H_1$ should be preferred so that a speaker turn is detected at time $i$.

### 2.2. Detection of all speaker turns

The distance $d_\mathrm{R}$ is computed for a pair of adjacent portions (windows) of the same size (about 2s), and the windows are then shifted by a fixed step (about 0.1s) along the whole parameterized speech signal. This process (see figure 2) gives as output the graph of distance with relation to time which is smoothed by a low-pass filtering operation. Since high values of $d_R$ correspond to speaker turns, all the “significant” local maxima are searched. A local maximum is regarded as “significant” when the differences between its value and those of the minima surrounding it are above a certain threshold (calculated as a fraction of the variance of the distance distribution), and when there is no higher local maximum in its vicinity. This detection method is detailed in [1]. Since a missed detection (an actual speaker turn has not been detected) is more severe for the verification process than a false alarm (a speaker turn has been detected although it does not exist), parameters involved in the speaker turn detection have been tuned to avoid missed detection to the detriment of false alarms. Thus, the parameterized signal is likely over-segmented (utterances of a given speaker are split into several segments).

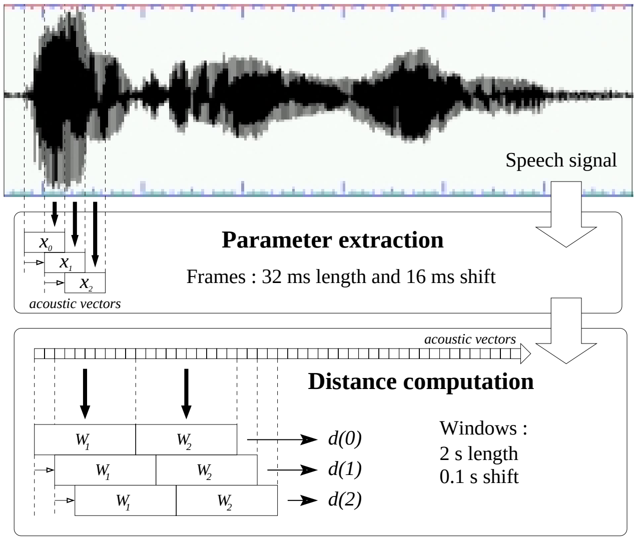

Figure 2: *Sliding windows*

## 3. SPEAKER VERIFICATION

Speaker verification has to be carried out on each of the segments resulting from the speaker turn detection, to detect which ones belong to the target speaker.

In order to minimize channel-related perturbations, Cepstral Mean Normalization (CMN) is applied to the feature vectors before the speaker verification process. The front-end segmentation allows this to be done on a segment-per-segment basis during the test phase, to reflect the fact that the various speakers may use different channels.

### 3.1. Speaker modeling

Speaker model training relies on the EM (Expectation-Maximization [2]) algorithm to estimate Gaussian Mixture Models (GMM [10]). Let $x$ be a $p$-dimensional feature vector of speech signal uttered by speaker $X_s$, the mixture density is defined as:

$$p(x|X_s) = \sum_{i=1}^{M} p_s^i \mathcal{N}(x, \mu_s^i, \Sigma s^i)$$

where $p_s^i$ and $\mathcal{N}(x, \mu_s^i, \Sigma s^i)$ are the mixture weights which satisfy constraint $\sum_{i=1}^{M} p_s^i = 1$ and the $i$-th unimodal gaussian density, summarized by mean vector $\mu_s^i$ and covariance matrix $\Sigma_s^i$.
In this paper, the gaussian mixtures are made of 16 components, for which full covariance matrices are used.

### 3.2. Similarity measure computation

This speaker verification system is based on a "block-segmental" approach ([5]). The speech signal is first split into short temporal blocks of fixed length (0.3 second) on which a similarity measure is computed (blocks located on segment boundaries are taken into account within the segment to which they belong most).

Normalization is applied on the similarity measures in order to cope with variability problems such as message content, noise and degradation due to signal recordings and transmission channels, and mismatch across training and testing conditions.

The normalization method combines two techniques classically used for speaker verification [4][5]:
• First, a classical world model-based likelihood ratio is computed for each block.
• Then, a Maximum A Posteriori (MAP) normalization is applied to the similarity ratios. Therefore, the normalized similarity measure for a block refers to the a posteriori probability of recognizing the target speaker.

Combination of the two techniques allows to learn a normalization function on a separate data set (of small size, thanks to the world model-based normalization), with no use of the target speaker model during the test phase.

### 3.3. Decision strategy

The final step of the speaker verification process consists in merging the block scores for each segment, to yield a segment score upon which to take a decision. A simple arithmetical mean is used here. Finally, in order to decide whether to attribute the segment to the target speaker, the merged score is compared to a threshold — which has been learned using a separate data set, and optimized for segment durations of 3 seconds.

## 4. EXPERIMENTS

### 4.1. Databases and parameterization

The data we used to assess our system consists in a subset of Switchboard II corpus used for the NIST/NSA 99 speaker verification campaign. This data set is made up of 230 male and 309 female speakers. The training material consists of two minutes of speech recorded over two sessions.

The various data sets used for system tuning are defined by the ELISA consortium [3]. The one used for the gender-dependent world model training is extracted from data of the NIST/NSA 98 speaker verification campaign. It is composed of recordings of 30 second long speech signal uttered by 100 male and 100 female speakers. Besides, a separate development data set made up of 100 male and 100 female speakers is used for the normalization function learning (see section 3.2).

Our STS has been experimented through 4000 trials of about one minute of speech signal each.

The speech signal has been represented, every 10ms, by 16 cepstrum coefficients derived from filter bank analysis (see [3]).

### 4.2. Assessment protocols

To assess our system in the framework of the NIST evaluations, two rates are computed: the false acceptance rate, which is the percentage of 10ms-blocks improperly attributed to the target speaker, and the false reject rate, which is the percentage of 10ms-blocks uttered by the target speaker and rejected by the system. These 10ms-blocks correspond to our acoustic vectors.

### 4.3. Results and comments

Figure 3 shows, in the form of DET curves, the results obtained by the various speaker tracking systems par-

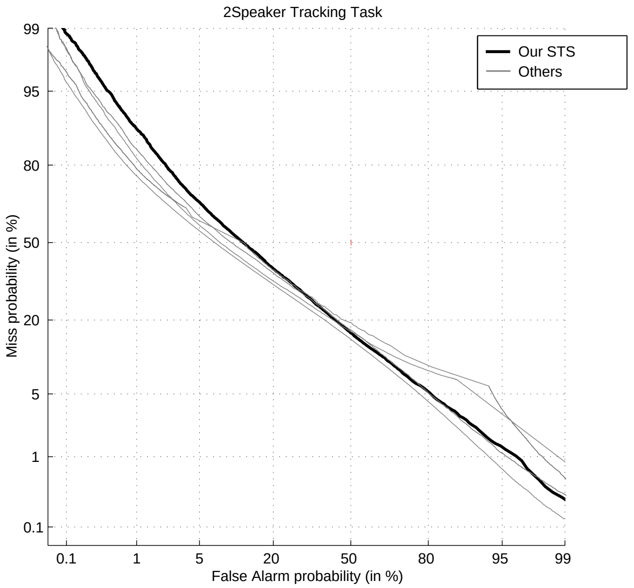

Figure 3: *DET curves of the various STS participating to the NIST/NSA 99 evaluation campaign.*

ticipating to the NIST/NSA 99 evaluation campaign.

The results obtained by our STS may be considered as correct, in view of the high complexity of the speaker tracking task and the low complexity of our architecture.

All the systems appear to have comparable performance. However, it has to be noticed that the assessment method does not favor the STS based on segment detection. Indeed, a shift of a segment boundary as minor as 10ms is considered as a false detection.

## 5. CONCLUSION

The proposed system makes use of two standard tools for speech signal analysis (speaker change detector and speaker verificator) and demonstrates good results at least equivalent to other more sophisticated systems at the NIST evaluation campaign. These encouraging results demonstrate the quality of both parts of the system which can also be used separately for building other applications. Several improvements are contemplated; first, the knowledge of a target speaker model can be taken into account at the segmentation level. Second, an improvement could be to turn local decision (on isolated segments) to a global decision on a set of segments that have been clustered at the end of the segmentation step.

Experiments should be conducted on multi-speaker tracking, where this system should perform quite well. A better adaptation to the specific conditions of the NIST campaign may also be realized by integrating the knowledge of the number of intervening speakers into the segmentation process.

This application, less traditional than simple speaker verification, contributes to the toolbox required to process speech signal as well as text or any other medium for multimedia document processing (indexing, editoring, semantic analysis,...).

## 6. REFERENCES

[1] P. Delacourt, D. Kryze, C.J. Wellekens, Speaker-based segmentation for audio data indexing, *ESCA workshop: accessing information in audio data*, 1999.

[2] D. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via EM algorithm, *J. Roy. Stat. Soc.*, Vol. 39, pp 1-38, 1977.

[3] ELISA consortium, The ELISA'99 Speaker Recognition and Tracking Systems, *Workshop on Automatic Identification Advanced Technologies*, Oct. 1999.

[4] C. Fredouille, J.-F. Bonastre, T. Merlin, Segmental normalization for robust speaker verification, *Workshop on robust methods for speech recognition in adverse conditions*, 1999.

[5] C. Fredouille, J.-F. Bonastre, T. Merlin, Similarity normalization method based on world model and a posteriori probability for speaker verification, *EUROSPEECH*, 1999.

[6] H. Gish, M.-H. Siu, R. Rohlicek, Segregation of speakers for speech recognition and speaker identification, *ICASSP*, pp 873-876, 1991.

[7] H. Gish, N. Schmidt, Text-independent speaker identification, *IEEE Signal Processing magazine*, pp 18-32, Oct. 1991.

[8] I. Magrin-Chagnolleau, A.E. Rosenberg, S. Parthasarathy, Detection of target speakers in audio databases, *ICASSP*, 1999.

[9] M.A. Przybocki, A.F. Martin, NIST Speaker Recognition Evaluation 1997, *RLA2C*, pp 120-123, Apr. 1998.

[10] D. A. Reynolds, Speaker identification and verification using gaussian mixture speaker models, *Speech Communication*, pp 91-108, Aug. 1995.