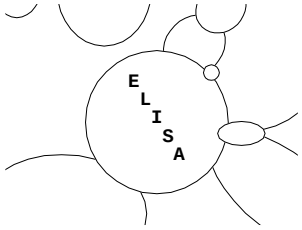
The ELISA Systems for the NIST'99 Evaluation in Speaker Detection and Tracking

The ELISA Consortium[1]

THE **ELISA** CONSORTIUM



The ELISA Consortium, The ELISA Systems for the NIST'99 Evaluation in Speaker Detection and Tracking, *Digital Signal Processing* **10** (2000), 143–153.

This article presents the text-independent speaker detection and tracking systems developed by the members of the ELISA Consortium for the NIST'99 speaker recognition evaluation campaign. ELISA is a consortium grouping researchers of several laboratories sharing software modules, resources and experimental protocols. Each system is briefly described, and comparative results on the NIST'99 evaluation tasks are discussed.

Key Words: text-independent, speaker verification, speaker detection, speaker tracking, NIST evaluation campaign

1. INTRODUCTION

The ELISA Consortium was founded in 1997 by a group of European laboratories, namely ENST (Ecole Nationale Supérieure des Télécommunications, Paris - France), EPFL (Ecole Polytechnique Fédérale de Lausanne - Switzerland), IDIAP (Institut Dalle Molle d'Intelligence Artificielle Perceptive, Martigny - Switzerland), IRISA (Institut de Recherche en Informatique et Systèmes

[1]The contributors to the ELISA Consortium, since its creation in 1997, have been (in alphabetical order) : Frédéric BIMBOT, Raphaël BLOUET, Jean-François BONASTRE, Gilles CALOZ, Jan ČERNOCKÝ, Gérard CHOLLET, Geoffrey DUROU, Corinne FREDOUILLE, Dominique GENOUD, Guillaume GRAVIER, Jean HENNEBERT, Jamal KHARROUBI, Ivan MAGRIN-CHAGNOLLEAU, Téva MERLIN, Chafic MOKBEL, Bojan NEDIC, Dijana PETROVSKA-DELACRETAZ, Stéphane PIGEON, Mouhamadou SECK, Patrick VERLINDE, Meriem ZOUHAL. E-mail: elisa@lia.univ-avignon.fr.

Digital Signal Processing Journal, Vol. 10, January, 2000, pp. 143-153

Aléatoires, Rennes - France) and LIA (Laboratoire d'Informatique d'Avignon - France), with the goal to build a common speaker recognition platform and participate to the NIST campaigns. The first participation took place in 1998. More recently, VUT (Vysoché Uçeni Technické, Brno - Czech Republic), RMA (Royal Military Academy, Brussels - Belgium), RICE University (Houston - USA) and FPMS (Faculté Polytechnique de Mons - Belgium) joined the Consortium and took part to the NIST'99 campaign.

The aim of the Consortium is to facilitate scientific communication and exchange in the field of text-independent speaker recognition and to share some of the development effort needed to participate, on a regular basis, in the NIST evaluation campaigns. A baseline modular software platform is maintained and regularly updated in a concerted manner within the Consortium, which tends to minimize useless duplication of software development. Nevertheless, the ELISA approach preserves individuality, as the partners are encouraged to submit to the NIST campaigns all kinds of variants of the ELISA baseline system, as primary or secondary systems, in order to evaluate the impact of a particular module of their own. ELISA also shares experimental protocols, which helps the partners to compare their approaches in between NIST campaigns.

Sections 2 and 3 of this article, describe briefly the speaker detection and tracking systems developed by different partners within the ELISA Consortium for the NIST'99 campaign [12]. The results obtained at the official evaluation are then analyzed in section 4. The systems are not described in full details since most of them are presented in articles specific to each site, published in this volume, which will be refered to whenever necessary [4, 6, 10, 14, 15, 20]. This article therefore concentrates on the contrastive comparison of the performances.

2. ONE-SPEAKER DETECTION

All ELISA one-speaker detection systems are based on a probabilistic framework using Gaussian Mixture Models (GMM) for speaker and non-speaker modeling, directly inspired from the state-of-the-art [17, 18, 19].

The ELISA systems can then be divided into two main categories (global *vs* segmental) : the global frame-based systems which model the distribution of the acoustic observations as a whole and the segmental systems which first proceed to some segmentation and pre-classification of the speech frames and then use specific distribution models for each class.

2.1. Global frame-based systems

Most of the ELISA'99 systems (ENST, IDIAP, IRISA, RIMO[3], and VERE[4]) use the GMM approach at the frame level. The differences between all these systems lie in :

- the type of acoustic analysis,
- the GMM estimation algorithm,
- the background modeling,
- the score normalization technique.

[3] **RI**ce university - Faculté Polytechnique de **MO**ns
[4] **V**UT - **E**PFL - **R**MA - **E**NST

The IRISA system is based on Maximum *A Posteriori* (MAP) estimation, applied to a 256 component GMM [5] with diagonal covariance matrices, and a gender-dependent background model. Log-likelihood ratios are computed for each frame, z-normalized [7] at the frame level [20], and averaged over all frames to give a final score.

The IDIAP, ENST and VERE systems are based on Maximum Likelihood (ML) trained GMM speaker models, with diagonal covariance matrices. The IDIAP system uses a 256 component model along with a handset-dependent background model together with a h-normalization (*i.e.* a handset-dependent z-normalization) [18], while the ENST system is based on a 128 component GMM, a gender- and handset-dependent background model together with h-normalization [6]. The VERE system is a multi-stream GMM with a 64 component GMM for the cepstral coefficient, a second 64 component GMM for the delta cepstral coefficients and a 16 component GMM for the delta energy. A gender-dependent background model is used and the z-normalization is applied to the likelihood ratio. Since the test segment length is not a priori known, VERE uses a first order polynomial approximation to compute the mean and standard deviation of the log likelihood ratio of the impostor scores from 3, 10 and 30 sec. test segments [14].

The RIMO system is also a 128 component GMM with diagonal covariance matrices, and uses a gender- and handset-independent background model. As mentioned above, the specificity of this system relies on an acoustic analysis which consists of time-frequency principal components computed from the output of a 24 channel Mel-frequency filter bank [9, 10]. All other ELISA systems use the classical cepstral front-end analysis with long-term mean subtraction (estimated over the whole utterance) and their first order derivatives.

A contrastive table (Table 1) summarizes the main differences between all the ELISA one-speaker systems

TABLE 1
Summary of the ELISA one-speaker systems

system	category	acoustic features	speaker model [a]	background model	normalization
ENST	global	16 LPCC+Δ+δE	128 – ML	gender+handset	h-norm
IDIAP	global	16 LPCC+Δ+δE	256 – ML	handset	h-norm
IRISA	global	16 MFCC+Δ	128 – MAP	gender	z-norm
LIA	segmental	16 MFCC	16[b] – ML	gender+handset	MAP
RIMO	global	TFPC	128 – ML	-	-
VERE	global	16 LPCC+Δ+δE	2×64+16[c] – ML	gender	z-norm

[a] number of components – training method
[b] with full covariance matrices
[c] number of components are given for each streams.

For all the systems, the operating thresholds are tuned experimentally on a development set composed of a subset of the NIST'98 evaluation data, which is a distinct speaker population from the NIST'99 evaluation data. A speaker-independent threshold is estimated in order to optimize the NIST'99 Decision Cost Function (see section 4.1) on the development data, and then this threshold is used for the evaluation.

2.2. Segmental systems

The LIA system (called AMIRAL) is based on partial scores calculated from speech segments rather than based on the global distribution of the acoustic frames. Fixed-length segments of 0.3 seconds are considered and the frame distribution for a segment is modeled by a 16 component GMM with full covariance matrices. The score for each segment consists of a handset and gender dependent likelihood ratio normalized using a MAP scheme that approximates the posterior probability of a speaker given the segment score [3]. The normalization function is learned on a subset of the NIST'98 evaluation data and does not require the computation of any impostor score distribution using the NIST'99 data. The segment scores are then averaged to give a final score on which the accept/reject decision is made, by comparison to a threshold.

The primary LIA system is a full-band system but multi-band variants of this system using dynamic information [2] rather than delta cepstral coefficients were also proposed. Results for these alternative systems are not reported here but can be found in [4].

As can be seen in Table 1, the LIA system is the only segmental system, within the ELISA Consortium, which actually took part to the NIST'99 evaluation. However, EPFL and VERE have also been working on segmental verification systems based on Automatic Language Independent Speech Processing (ALISP) techniques [1, 13]. The speech segmentation is obtained using Temporal Decomposition (TD) followed by a quantization into 8 classes of the TD spectral targets. GMM's are used for segmental speaker modeling. The client score for a frame is calculated as a weighted sum of the class-dependent GMM scores, the weights accounting for the probability that the frame belongs to a given class. This approach was evaluated on a subset of the NIST'98 data and was judged disappointing [21, 14] as compared to the segmental approaches based on Multi-Layer Perceptron (MLP) speaker modeling [13].

2.3. Fusion experiments

A fusion of the results of the different systems described above was performed by RMA, using logistic regression, a method which linearly combines the outputs of the individual systems (experts) and maximizes a likelihood function based on the logistic regression model.

Logistic regressions have been successfully used in the past for fusing several image and speech experts together [22]. The approach is based on the assumption that the individual experts are independent. Such an assumption is probably not satisfied by the ELISA systems used as experts and their correlation can yield a lower expectancy of the fusion gain.

In the experiments reported in this article, the coefficients of the logistic regression are learned on the male test population and are then used to combine the scores of the various systems on the female test population. More comprehensive results are given in [15].

3. TWO-SPEAKER DETECTION AND SPEAKER TRACKING

3.1. Two-speaker detection

IRISA and LIA extended their one-speaker systems in order to perform the two-speaker detection task. Techniques similar to the ones used for the one-speaker task are used to detect a speaker in a conversation. However, the following features are specifically integrated for the two-speaker detection task :

- the LIA system computes the utterance score only from a subset of frames, taking into account the histogram of the frame-based likelihood values and performs short-term cepstral mean subtraction (using 3 second windows) in order to adapt this channel compensation technique to the fact that the 2 handsets have distinct transfer functions [4],
- the IRISA system uses ML estimates of the client models rather than MAP estimates, as was the case for the one-speaker task, and does not use z-normalisation [20].

In an experiment carried out after the end of the official campaign, IRISA modified its system (refered to as IRISA2 in this article), including a step that first estimates the proportion of the target speaker's speech (based on the ML mixture estimation of the target and background GMM models) and then uses this estimate to score the utterance against the target speaker [20].

3.2. Speaker tracking

The ELISA speaker tracking systems are also based on frame scoring and use additional smoothing techniques to obtain less noisy scores along the speech utterance. Segmentation algorithms are then applied to locate the beginning and the end of segments corresponding to the target speaker. IRISA, LIA, and RIMO provided such systems.

The IRISA tracking system uses a smoothed log-likelihood ratio calculated on blocks of 5 frames. The decision is taken for each block independently [20].

The LIA tracking system uses fixed-length blocks of frames and a two-pass algorithm. During an initial pass, a score based on a likelihood ratio is calculated on each block and a subset of these blocks are selected according to a statistical criterion on the distribution of their score. A decision is then made on the presence of the target speaker by comparing the average score of the selected blocks to a first threshold. If the target speaker is detected, a second pass is carried out to label each block as target or non-target, by comparison of the score to a second threshold [4].

The RIMO speaker tracking system is based on TFPC analysis [10] and on a sequential segmentation algorithm using multiple thresholds, inspired from [8].

4. RESULTS AND DISCUSSION

4.1. Experimental setup

The NIST'99 evaluation corpus was extracted from the Switchboard II - Phase 3 corpus. The former contains 539 speakers (309 females and 230 males) and two sessions of about 1 minute each were provided to estimate the speaker model parameters. The two training sessions were choosen so that the handset microphone type is the same for the two telephone calls.

Each task is assessed using a detection cost function (DCF) given by :

$$\mathrm{DCF} = C_{\mathrm{fr}}\, P_{\mathrm{target}}\, P_{\mathrm{fr}} + C_{\mathrm{fa}}\, P_{\overline{\mathrm{target}}}\, P_{\mathrm{fa}}$$

where C_{fr} (resp. C_{fa}) is the cost of a false rejection (resp. of a false acceptance) and P_{target} (resp. $P_{\overline{\mathrm{target}}}$) is the prior probability of a genuine speaker trial (resp. an impostor trial). The costs were set to $C_{\mathrm{fr}} = 10$ and $C_{\mathrm{fa}} = 1$ while the prior probabilties were $P_{\mathrm{target}} = 0.01$ and $P_{\overline{\mathrm{target}}} = 0.99$.

P_{fr} and P_{fa} are the measured false rejection and false acceptance rates which, for the speaker tracking task, were defined as :

$$P_{\mathrm{fr}} = \frac{\text{number of true target frames labeled as non target}}{\text{number of target frames}}$$

$$P_{\mathrm{fa}} = \frac{\text{number of non target frames labeled as true target}}{\text{number of non target frames}}$$

In all cases, scores were also provided with each binary decision to compute the detection error tradeoff (DET) curves [11], showing how false rejections may be traded off against false acceptances as a function of the decision threshold.

A more detailed presentation of the database and of the evaluation measures can be found in [12].

4.2. One-speaker detection

The DET curves obtained by the primary systems of the members of the ELISA Consortium, for the one-speaker detection task are given in Figure 1. The plus signs on the DET curves indicate the operating points for which the DCF is minimal, while the circles indicate the actual operating points, *i.e.* the DCF corresponding to the decisions that were actually made. Since the first evaluation campaigns pointed out the importance of channel (*i.e.* telephone lines) and handset-type (*i.e.* carbon button *vs* electret microphones) normalization, the DET curves present separately three test conditions. The first condition groups together the segments for which the number of the caller and the handset type are the same between the target speaker training segments and the test segment

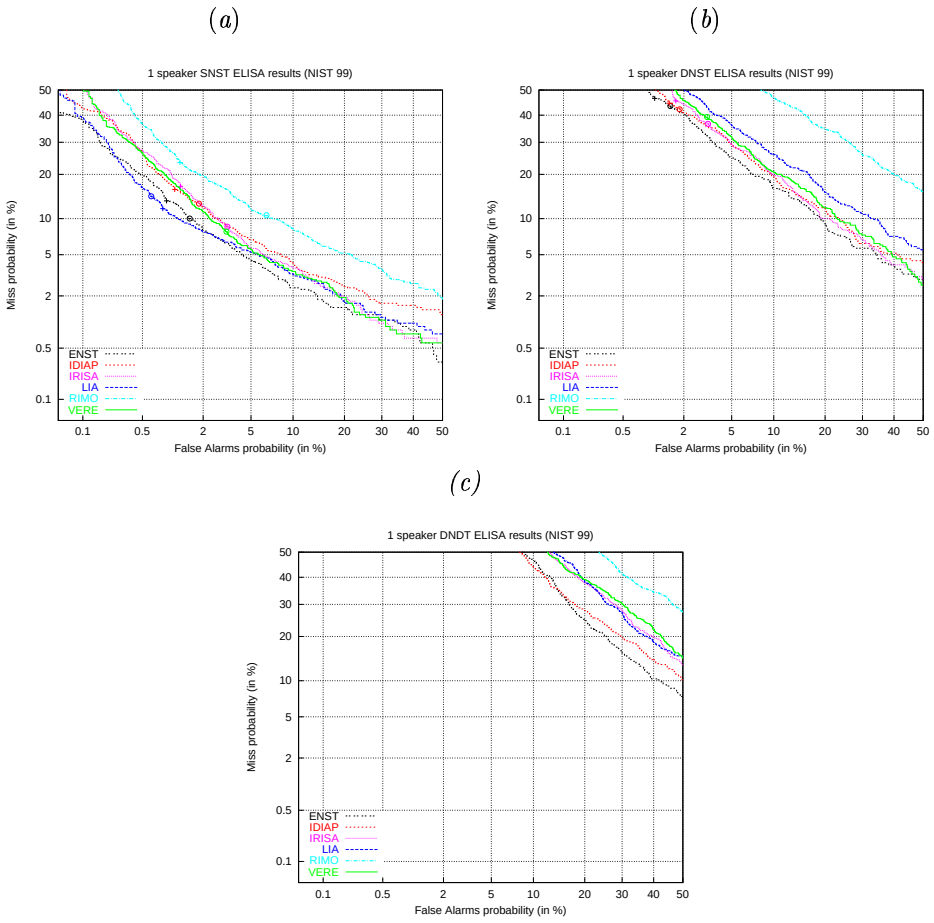


FIG. 1. DET curves of the ELISA one-speaker systems, for different poolings : (*a*) SNST, (*b*) DNST, (*c*) DNDT.

(SNST for *Same Number Same Type*)[5]. For the second condition, training segments and test segments are from different telephone lines but share the same handset type (DNST for *Different Number Same Type*). Finally, both the channel and the handset type are different in the third condition (DNDT for *Different Number Different Type*). In Figure 1, the (*a*), (*b*) and (*c*) curves respectively correspond to the first, second and third condition.

Several conclusions can be drawn from these results. It can be seen that in the SNST match condition (curve (*a*)), most of the systems show comparable performances. While the ENST and LIA systems show a slight advantage over the others, the RIMO system performs significantly worse, which is probably due

[5]It must be stressed that the same number condition only applies to target speaker trials as it never occurs, in the evaluation data, that an impostor uses the same telephone as the target speaker.

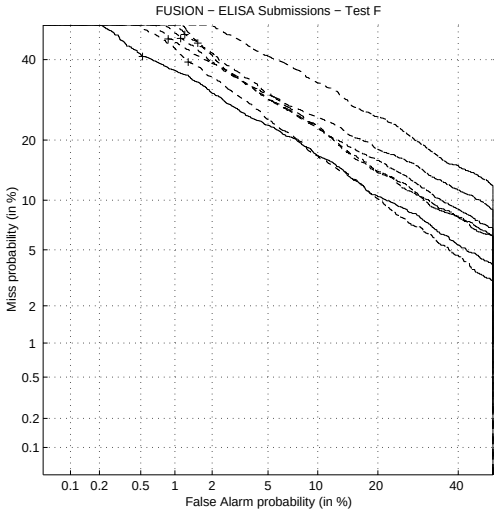


FIG. 2. DET curve for the fusion of the ELISA one-speaker systems (female only). The solid line corresponds to the fusion DET curve while the dotted ones correspond to the individual systems.

to an improper normalization as a result of the two likelihood functions operating in different acoustic spaces.

For the DNST mismatch condition (curve (b)), the global performance decreases significantly for all systems. The performance of the IDIAP, VERE and IRISA systems are getting closer to the ENST system. On the contrary, the LIA system performance degrades more than for the other systems. One explanation is that the delta coefficients are not used in the LIA system and it is plausible that the delta coefficients are more robust to channel mismatch. Therefore they are more useful in the DNST condition, and their absence penalises the LIA system for that condition.

Finally, in the DNDT condition (curve (c)), the curves can roughly be pooled into three groups. The main difference between the first group and the other systems probably comes from the use of handset knowledge for the selection of the background model and in the likelihood ratio normalization scheme. A study on the influence of prior knowledge for normalization can be found in [6]. It can also be seen that the performance of the LIA system is now similar to the IRISA and VERE systems which was not the case for the DNST condition. This suggests that the LIA system is more sensitive to channel mismatch than to handset microphone mismatch.

The fusion results, depicted in Figure 2 show that the use of logistic regression to combine the results of individual systems brings a clear benefit compared to the performance of the best individual expert, with a 50% reduction of the false acceptance, when keeping a false rejection level around 40%.

4.3. Speaker detection and tracking

Figure 3 (a) shows the DET curves for the official LIA and IRISA two-speaker detection system, and the DET curve of the non-official IRISA system (labeled as

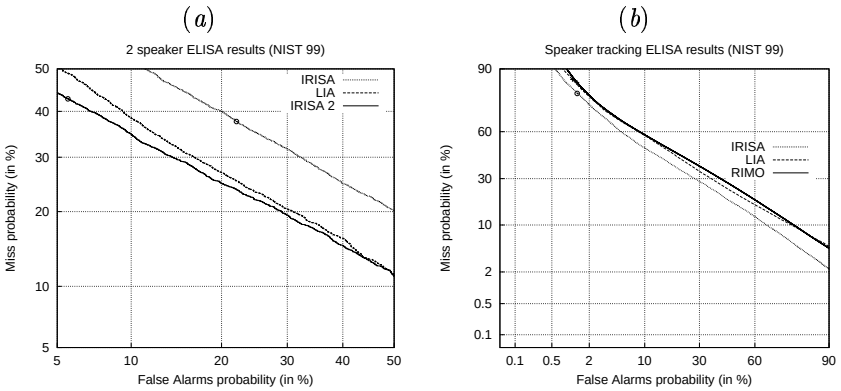


FIG. 3. ELISA DET curves for the two-speaker detection (a) and the speaker tracking (b) tasks.

IRISA 2). Figure 3 (b) depicts the DET curves for the speaker tracking systems in the DNST condition, where both sides of the conversation are electret and both speakers are of the same gender.

The LIA two-speaker tracking system clearly outperforms the official IRISA system. This can certainly be explained by the fact that the likelihood score of the IRISA system is computed over the whole utterance and therefore biased by the second speaker's speech, whereas the LIA system first selects a subset of frames according to their target speaker likelihood, and scores the utterance only on that subset. This origin of the weakness of the official IRISA system is confirmed by the gain obtained with the non-official system, which better models the fact that the utterance is composed of two speakers.

For the speaker tracking task, the IRISA system tends to outperform the other ELISA systems.

5. CONCLUSION

In this article, several ELISA variants of one-speaker detection, two-speaker detection and speaker tracking systems, were presented and compared on the NIST'99 evaluation data.

The one-speaker detection results indicate that the most significant differences in performance within these systems come from the normalization techniques. The design of the speaker independent background model seems to be essential, the second important factor being the likelihood ratio normalization scheme. In particular, it is clear that the appropriate use of handset type knowledge is beneficial to the performance.

The ELISA platform for one-speaker detection was extended in several directions in order to perform the two-speaker detection and speaker tracking tasks. The NIST'99 campaign has helped identifying a number of factors that influence the performance for these two new tasks, but efforts are required to consolidate these initial trends.

6. PROPOSITIONS AND PERSPECTIVES

For the forthcoming evaluations, the ELISA Consortium has put forward a few proposals :

- to introduce a fraction of non (american) english data, so that language dependence and cross-language mismatch can be studied, as well as the impact of language mismatch on LVCSR based systems,

- to define several DCF functions in the evaluation plan, and to disclose only after the results are returned by the participants, which of the DCF is the primary one; this would prevent systems from becoming tuned to a particular operating point, and certainly stimulate research on the issue of decision threshold setting,

- to release the keys for the evaluation database in two steps, corresponding to two distinct subsets of target speakers, so that some participants can return, between the two releases, fusion results on the second part of the eval data base obtained from systems trained on the first part.

In the long-term, the ELISA Consortium intends to continue participating on a regular basis in the NIST speaker recognition evaluation campaigns. Its multi-site structure is a favorable factor for scientific progress and against fluctuations in local conjonctures. However, such a working structure also requires specific procedures for collective software engineering and internal quality control.

REFERENCES

1. G. Chollet, J. Černocký, G. Gravier, J. Hennebert, D. Petrovska-Delacrétaz, and F. Yvon. Towards fully automatic speech processing techniques for interactive voice servers. In G. Chollet, M. Di Benedetto, A. Esposito, and M. Marinaro, editors, *Speech Processing, Recognition and Artificial Neural Networks. Proceedings of the 3rd International School on Neural Nets "Eduardo R. Caianiello"*. pp. 297–326, Springer-Verlag, 1998.
2. C. Fredouille and J.-F. Bonastre. Use of dynamic information with second order statistical methods in speaker identification. In *Proc. of the Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pp. 50–54, Avignon, 1998.
3. C. Fredouille, J.-F. Bonastre, and T. Merlin. Similarity normalization method based on world model and a posteriori probability for speaker verification. In *Proc. Eurospeech'99*, pp. 983–986, 1999.
4. Corinne Fredouille, Jean-Francois Bonastre, and Teva Merlin. AMIRAL: a bloc-segmental multi-recognizer architecture for automatic speaker recognition. *Digital Signal Processing*, 10(1), 2000 (*this volume*). Academic Press.
5. Jean-Luc Gauvain and Chi-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2), pp. 291–298, April 1994.
6. Guillaume Gravier, Jamal Kharroubi, and Gérard Chollet. On the use of prior knowledges in normalization schemes for speaker verification. *Digital Signal Processing*, 10(1), 2000 (*this volume*). Academic Press.
7. Kung-Pu Li and Jack E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *Proc. ICASSP'88*, pp. 595–597, 1988.
8. Ivan Magrin-Chagnolleau, Aaron E. Rosenberg, and S. Parthasarathy. Detection of target speakers in audio databases. In *Proc. ICASSP'99*, pp. 821–824, 1999.
9. Ivan Magrin-Chagnolleau and Geoffrey Durou. Time-frequency principal components of speech: Application to speaker identification. In *Proc. Eurospeech 99*, pp. 759–762, September 1999.
10. Ivan Magrin-Chagnolleau and Geoffrey Durou. Application of time-frequency principal component analysis to speaker verification. *Digital Signal Processing*, 10(1), 2000 (*this volume*). Academic Press.

11. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech'97*, pp. 1895–1898, 1997.

12. Alvin Martin and Mark Przybocki. The NIST 1999 Speaker Recognition Evaluation - An overview. *Digital Signal Processing*, 10(1), 2000 (*this volume*). Academic Press.

13. D. Petrovska-Delacrétaz, J. Černocký, J. Hennebert and G. Chollet. Text-Independent Speaker Verification using Automatically Labelled Acoustic Segments. In *Proc. ICSLP'98*, Paper # 536, 1998.

14. Dijana Petrovska-Delacrétaz, Jan Černocký, Jean Hennebert, Gérard Chollet. Segmental Approaches for Automatic Speaker Verification. *Digital Signal Processing*, 10(1), 2000 (*this volume*). Academic Press.

15. Stéphane Pigeon, Pascal Druyts, and Patrick Verlinde. Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing*, 10(1), 2000 (*this volume*). Academic Press.

16. Mark A. Przybocki and Alvin F. Martin. NIST Speaker Recognition Evaluation - 1997. In *Proc. of the Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pp. 120-123, Avignon, 1998.

17. Douglas A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. In *Proc. of the ESCA Workshop on Speaker Recognition, Identification and Verification*, pp. 27–30, Martigny, 1994.

18. Douglas A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proc. Eurospeech'97*, pp. 963–966, 1997.

19. Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1), 2000 (*this volume*). Academic Press.

20. Mouhamadou Seck, Raphaël Blouet, and Frédéric Bimbot. The IRISA / ELISA speaker detection and tracking systems for the NIST'99 evaluations campaign. *Digital Signal Processing*, 10(1), 2000 (*this volume*). Academic Press.

21. J. Černocký, D. Petrovska-Delacrétaz, S. Pigeon, P. Verlinde, and G. Chollet. A segmental approach to text-independent speaker verification. In *Proc. Eurospeech'99*, pp. 2207–2210, 1999.

22. P. Verlinde and G. Chollet. Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application. In *Proc. of the Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA)*, pp. 188-193, Washington D.C., 1999.