The ELISA'99 Speaker Recognition and Tracking Systems

G. Gravier, J. Kharroubi and G. Chollet (ENST-TSI and CNRS-URA 820)
D. Petrovska (EPFL), G. Durou (FPMS), B. Nedic (IDIAP)
F. Bimbot, R. Blouet and M. Seck (IRISA)
J.-F. Bonastre, C. Fredouille and T. Merlin (LIA)
I. Magrin-Chagnolleau (Rice University)
S. Pigeon and P. Verlinde (RMA), J. Černocký (VUT)

elisa-consortium@onelist.com

Abstract

This article presents the text-independent speaker verification and tracking systems developed by the ELISA consortium for the NIST'99 speaker recognition evaluation campaign. ELISA is a consortium grouping European researchers of several laboratories sharing resources and experimental protocols. Each system is briefly described, and comparative results on the evaluation corpus are given.

I The NIST Evaluation Task

A Context

Since 1996, the National Institute of Standards and Technologies (NIST) organizes benchmark evaluations in text-independent speaker recognition over the telephone [1][1]. After distributing development data, NIST provides the participants with some unlabeled test data and each participant must return the decision of its system within a given time frame. NIST then assesses the different systems and distributes the results to the participants. Both academic and industrial laboratories take part each year in the evaluations whose main objectives are to experiment new ideas in the field of text-independent speaker recognition, to develop and implement the corresponding techniques and to assess state of the art and emerging techniques. These yearly evaluations are carried out on subsets of the Switchboard corpus.

[1] See also http://www.itl.nist.gov/iaui/894.01/test.htm for further information.

B Evaluation conditions

1) Task descriptions

The 1999 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 evaluation is carried out on the three following tasks

Speaker detection: this task is the conventional speaker verification task in which one must determine whether a specified target speaker is speaking during a given speech segment or not.

Speaker detection in a conversation: this task is essentially the same as the previous one, except that the test segments contain both sides of a telephone call, rather than being limited to the speech from a single speaker.

Speaker tracking: again for this task, the test segments contain speech from two speakers and the goal is to determine where a specific speaker is speaking in the conversation.

2) Assessment

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

$$\mathrm{DCF} = C_{\mathrm{fr}}\, P_{\mathrm{target}}\, P_{\mathrm{fr}} + C_{\mathrm{fa}}\, P_{\overline{\mathrm{target}}}\, P_{\mathrm{fa}} \ , \qquad (1)$$

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). For the 99 evaluation, the costs were set to $C_{\text{fr}} = 10$ and $C_{\text{fa}} = 1$ while the prior probabilties were $P_{\text{target}} = 0.01$ and $P_{\overline{\text{target}}} = 0.99$.

In (1), P_{fr} and P_{fa} are the measured false rejection and false acceptance rates. For the speaker tracking tasks, these rates are defined by

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

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

In all cases, scores were also provided with each binary decision to compute the detection error tradeoff (DET) curves [2], showing how false rejections may be traded off against false acceptances as a function of the decision threshold.

II The ELISA Systems

The ELISA consortium was created by ENST, LIA, and IRISA, along with EPFL and IDIAP, to participate in the NIST'98 campaign. More recently, VUT, RMA, Rice University, and FPMS joined the consortium. Although most laboratories have their own system, some teamed up to develop a common system. The members of the consortium share software resources and implementation issues, thus allowing to run many systems under the same test conditions in order to compare variants of the same baseline system.

A Speaker Verification

Most of the systems use the classical cepstral front-end analysis with long-term mean subtraction and are based on Gaussian Mixture Models (GMM) [3] with diagonal covariance matrices. The ELISA systems can be divided into two main categories. The first one gathers the classical text-independent systems for which a statistical model, the GMM, is given for the global frame distribution, while the second one groups together segmental systems for which the frame distribution is considered on a segment rather than globally.

1) Classical Approaches

Most of the ELISA'99 systems (ENST, IDIAP, IRISA, RIMO[2], and VERE[3]) used the classical GMM approach. The differences between all the systems mainly come from the choice of the training algorithms, the background modeling, and the score normalization techniques. The RIMO system also differs by its acoustic processing.

The IRISA system is based on maximum a posteriori trained 256 component GMM [4] and a gender-dependent background model. The log-likelihood ratio is computed for every frame, z-normalized [5] and averaged over all frames.

The IDIAP, ENST and VERE systems are based on maximum likelihood trained speaker models. The IDIAP system uses a 256 component model along with a handset dependent background model and h-normalization [5], while the ENST system is based on a 128 component mixture model, a gender and handset dependent background model and h-normalization. The VERE system is a multi-stream GMM with two 64 component mixtures for the static and delta cepstral coefficients and a 16 component mixture for the delta energy, and with a gender-dependent background model. The decision is made after z-normalization of the likelihood ratio.

The RIMO system also is a 128 component GMM and implements a gender and handset independent background model. The originality of this system relies on the acoustic analysis which consists in time-frequency principal components computed from the output of a 24-channel Mel-frequency filter bank [6].

2) Segmental Approaches

The LIA AMIRAL system is based on speech segments rather than on the global distribution of the feature vectors. It operates on fixed-length segments of 0.3 sec. and is based on 16 component mixture models with full covariance matrix. The score for each segment is a handset and gender dependent likelihood ratio normalized using a MAP scheme to give the posterior probability of recognizing a speaker knowing the segment score [7]. The segment scores are then averaged to give a final score.

The primary LIA system was a full-band system but multi-band variants of this system using dynamic information [8] rather than delta cepstral coefficients were also proposed. Results for these alternative systems are not reported here.

3) Fusion of the system decisions

The fusion of the different systems described above was performed by the RMA using logistic regression, a method which linearly combines the outputs of the individual systems and maximizes a

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

likelihood function based on the logistic regression model [9].

B Speaker Detection in a Conversation

This task relies on the same techniques as the previous one with the difference that the test segments must be pruned before making the decision. The test segments contain a conversation between two speakers and at least one of them is not the target speaker. Scores are therefore computed for each frame or group of frames using the target speaker model and the background models. Frames for which the score falls below a threshold are discarded, the remaining ones being kept and used as the new test segment on which to perform the detection. The IRISA and LIA systems were extended for this task.

C Speaker Tracking Systems

The ELISA speaker tracking systems were also based on frame scoring, and used additional smoothing techniques to obtain a block-based score. Various segmentation algorithms were then applied to locate beginning and ending time of segments corresponding to the speaker who was tracked. IRISA, LIA, and RIMO provided such systems.

The IRISA tracking system used a smoothed log-likelihood ratio calculated on non-overlapping blocks of 5 frames. The decision was then made for each block. The LIA tracking system used also fixed-length blocks of frames and an algorithm in two passes. During the first pass, a score based on a likelihood ratio was calculated on each block and compared to a first threshold. During the second pass, a score based on a weighted geometric mean was calculated, and compared to a second threshold. The RIMO tracking system was based on a sequential segmentation algorithm using multiple thresholds [10].

Results for the tracking systems are not presented in the paper.

III Results

The DET curves for the speaker detection task are given in Fig. 1, the plus signs indicating 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. It can be seen that most of the systems using cepstral analysis have similar performances. The different variants of the GMM

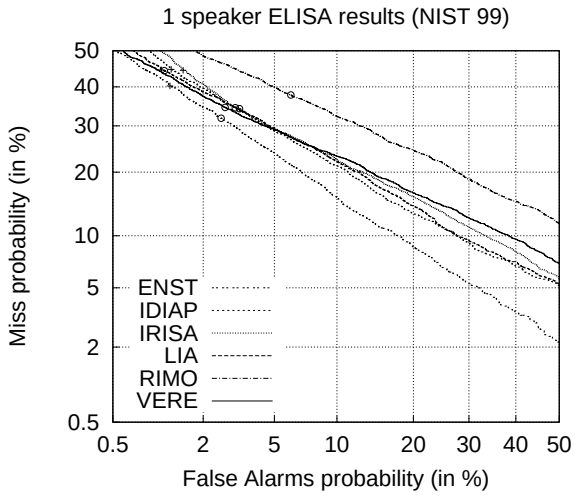

Fig. 1: DET curve for the speaker detection task.

and the different normalization techniques used in the systems do not make a real difference. A more detailed study of the results showed that some systems are more robust to channel mismatch than others. For example, the LIA system performs better than the ENST one when the same telephone line is used for training and testing, while the opposite is true when different telephone lines are used. Another interesting point to note is that the LIA system does not use any delta cepstral coefficients and still performs well.

In Fig. 2, the result of the fusion of the decision of all the systems is represented, the solid line corresponding to the fusion while the dotted ones represent the experts. Since it is necessary to estimate the fusion parameters, results are given only for female speakers with fusion parameters estimated on the male speakers. Fusion of highly correlated decisions still gives slightly better results than the best expert, specially around the minimum DCF points. This comes from the fact that the logistic regression parameters were estimated to minimize the DCF.

The results for the speaker detection in a conversation task are given in Fig. 3.

IV Conclusions and Perspectives

State of the art speaker recognition, detection and tracking systems developed within the ELISA consortium are presented in the paper, and results obtained in the NIST'99 speaker recognition campaign are presented. It was illustrated that most of the systems have similar performances and that

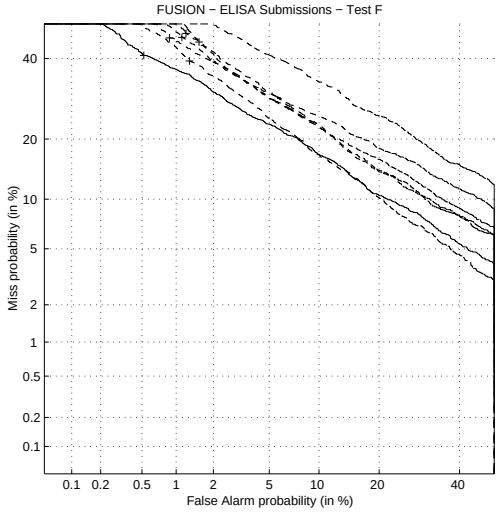

Fig. 2: DET curve for the fusion results (solid line) in the one speaker task (female speakers only). Dotted line represents the different system DET curves.

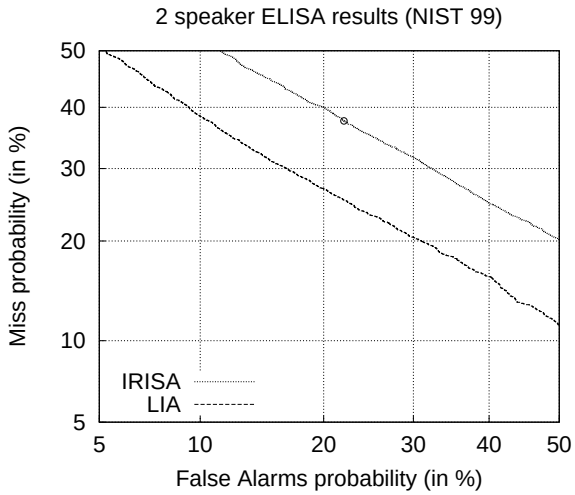

Fig. 3: DET curve for the speaker detection in a conversation task.

the fusion of even highly correlated decisions can improve the results. However, a more detailed study of the different methods should lead to the development of new algorithms for speaker recognition and to a better understanding of the weaknesses and the strengths of the methods presented here.

References

[1] Mark A. Przybocki and Alvin F. Martin. NIST speaker recognition evaluation - 1997. In *Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, 1998.

[2] A. Martin et al. The DET curve in assessment of detection task performance. In *Eurospeech*, volume 4, pages 1895–1898, 1997.

[3] Douglas A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. In *ESCA Workshop on Speaker Recognition, Identification and Verification*, pages 27–30, 1994.

[4] 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), April 1994.

[5] G. Gravier and G. Chollet. Comparison of normalization techniques for speaker verification. In *Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 97–100, 1998.

[6] Geoffrey Durou and Ivan Magrin-Chagnolleau. Time-frequency principal components of speech: Application to speaker identification. In *To be published in Eurospeech*, Sept. 1999.

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

[8] C. Fredouille and J.-F. Bonastre. Use of dynamic information with second order statistical methods in speaker identification. In *Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, 1998.

[9] 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 *Second Intl. Conf. on Audio-and Video-based Biometric Person Authentication (AVBPA)*, March 1999.

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