Overview of the 2000-2001 ELISA consortium research activities

Ivan Magrin-Chagnolleau, Guillaume Gravier, Raphaël Blouet

OVERVIEW OF THE 2000-2001 ELISA CONSORTIUM RESEARCH ACTIVITIES

Ivan Magrin-Chagnolleau, Guillaume Gravier, and Raphaël Blouet

for the ELISA consortium.

elisa@listes.univ-avignon.fr

ABSTRACT

This paper summarizes the research activities in speaker recognition in the framework of the ELISA consortium. The ELISA speaker recognition common platform is first presented, including the common evaluation protocol and the functioning of the consortium. Then experiments with this platform on the development data of the NIST 2001 speaker recognition campaign are reported. Finally, a survey of the research directions in the various ELISA laboratories is given.

1. INTRODUCTION

The ELISA consortium was originally created by ENST, EPFL, IDIAP, IRISA and LIA in 1998 with the aim of developping a common state-of-the-art speaker verification system and participating in the yearly NIST speaker recognition evaluation campaigns. Along the years, the composition of the consortium has changed and today ENST, IDIAP, IRISA and LIA are members of the consortium. Since 1998, the members of the Consortium have been participating in the NIST evaluation campaigns in speaker recognition and a comparative study of the various systems presented in the 1999 campaign can be found in [2].

The aim of the Consortium is to promote scientific exchanges between members. To reach this goal, a common baseline reference platform is maintained by all the members. The reference platform is modular in order to be easily modified and reflects state-of-the-art performance achieved with gaussian mixture models (GMMs). Modules are provided for the various tasks of the NIST evaluations, namely speaker verification, detection, tracking, and segmentation. The possibility of the platform to deal with segmental approaches [4] at the score computation level enables an easy integration of the speaker detection, tracking, and segmentation tasks in the same platform. A common evaluation protocol, derived from the NIST evaluation rules, is shared by all the Consortium members to allow fair comparisons between the variants of the baseline system.

In this paper, we describe in Section 2 the common resources of the ELISA consortium, including the architecture of the platform, the common evaluation protocol, and the functioning of the consortium. In Section 3, we report on some experiments that were carried out to bring the platform to state-of-the-art performance in speaker verification. In Section 4, we point out the various research directions of the laboratories member of the Consortium.

2. THE ELISA COMMON RESOURCES

2.1. Platform Architecture

The ELISA platform is composed of the following main modules: *speech parameterization*, *modeling*, *likelihood calculation*, *normalization*, and *decision / scoring*. The overall architecture of the platform is illustrated on Fig. 1. The *speech parameterization* module implements classical speech analyses such as filterbank analysis or cepstral analysis plus frame selection methods. The *modeling* module is based on gaussian mixture models (GMMs) with maximum likelihood (ML) parameter estimation and/or maximum a posteriori (MAP) adaptation of a speaker independent model. Various MAP adaptation techniques have been implemented. The *likelihood* module is in charge of log-likelihood computation and of mixture component selection for scoring. Several likelihood ratio normalization procedures are possible in the *normalization* module, such as z-norm [8], h-norm [13], t-norm [1], and World + MAP [5]. The *decision / scoring* module makes the decision by comparing a normalized likelihood ratio to a threshold and plots the DET curves [10].

2.2. Common Evaluation Protocol

The main goal of the ELISA Common Evaluation Protocol (CEP) is to make the results comparable within the Consortium while enabling the preparation of the next NIST

The list of the current members of the consortium is, in alphabetical order: F. Bimbot, R. Blouet, J.F. Bonastre, G. Chollet, C. Fredouille, G. Gravier, J. Kharroubi, I. Magrin-Chagnolleau, J. Mariethoz, S. Meignier, T. Merlin, and M. Seck.

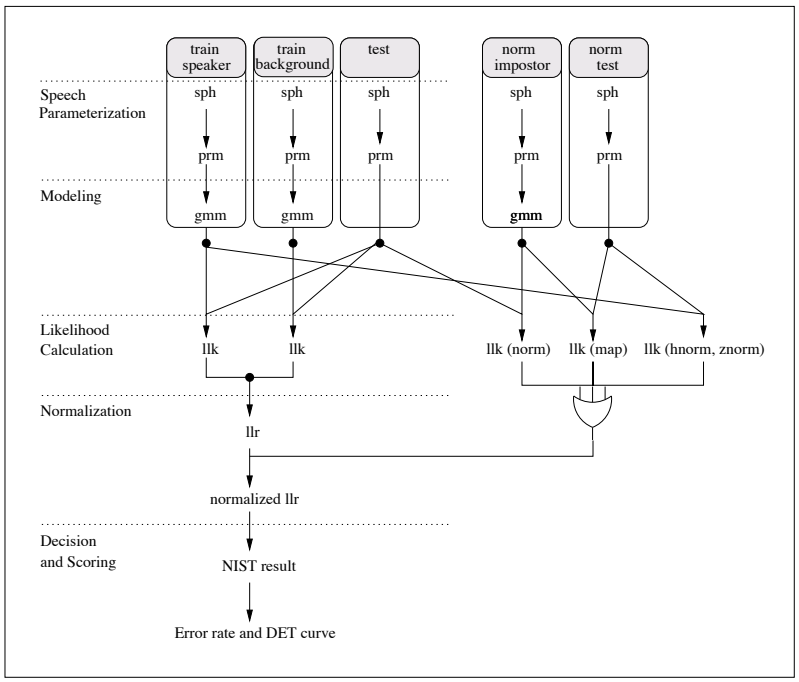

Fig. 1. *Architecture of the ELISA reference platform.*

evaluation campaign. The protocol is therefore redefined every year on a subset of the development data of the upcoming NIST evaluation campaign. The CEP for the 2001 campaign was defined as follows. 4 subsets of the NIST development data were defined. The first subset, called the *world* subset, is composed of 316 speakers (186 females using electret handsets, 10 females using carbon handsets, 102 males using electret handsets, and 18 males using carbon handsets), and is used to train gender and handset dependent world models. The second subset, called the *dev* subset, is composed of 100 speakers (50 females and 50 males), and is used to do some development experiments. The third subset, called the *eva* subset, is also composed of 100 speakers (50 females and 50 males), and is used to cross-validate the experiments done on the *dev* subset. Finally, the fourth subset, called the *norm* subset, is composed of 159 speakers (50 females using electret handsets, 24 females using carbon handsets, 50 males using electret handsets, and 35 males using carbon handsets), and is used for various normalizations of the log-likelihood ratios. These four subsets are used by all the members of the Consortium in order to have comparable results.

2.3. Functioning of the Consortium

Members of the consortium have regular meetings during the year. In average, they meet once every trimester. These meetings are the occasion to discuss current development and research issues, to compare the last results obtained by each laboratory, and to set up new goals until the next meeting. Each laboratory decides to focus on a particular aspect in order to avoid redundant work.

Each year, a standard configuration is defined as being the baseline configuration. A new configuration is integrated only if it has been proven to provide better performance than the baseline configuration. In that case, the new configuration becomes the new baseline system.

Additionally, results are regularly shared on the web site of the consortium, and a mailing list allows the consortium members to regularly discuss encountered problems, or any topic of interest.

3. EXPERIMENTS

After a presentation of the common resources of the ELISA consortium and its functioning, we present in this section experiments carried out to bring the platform to state-of-the-art performance.

3.1. Task

Although the ELISA platform can be used to deal with any of the tasks proposed in the NIST speaker recognition evaluation, namely speaker verification, speaker detection, speaker tracking, or speaker segmentation, we focus in this section on text-independent speaker verification (called *one speaker*

detection by NIST), which consists in verifying a claimed identity from a recorded speech utterance, without using any prior phonetic knowledge.

3.2. Database

Experiments are reported on the *dev* subset described in the common evaluation protocol.

3.3. Speech Analysis

Each speech utterance is converted from a μ-law into a linear representation with a sampling frequency of 8 kHz. Each utterance is decomposed into frames of 20 ms extracted every 10 ms. A Hamming window is applied to each frame. The signal is not pre-emphasized. For each frame, a fast Fourier transform is computed and provides square modulus values representing the short term power spectrum in the 0-4000 Hz band. This Fourier power spectrum is then used to compute 24 filterbank coefficients, using triangular filters placed on a linear frequency scale in the 300-3400 Hz band. The base 10 logarithm of each filter output is taken and multiplied by 10, to form a 24-dimensional vector of filterbank coefficients in dB. Then, cepstral coefficients [12] c_1 to c_{16}, augmented by their Δ coefficients [6] (calculated over 5 vectors) are calculated, and a cepstral mean subtraction (CMS) is applied. We finally obtain 32-dimensional feature vectors.

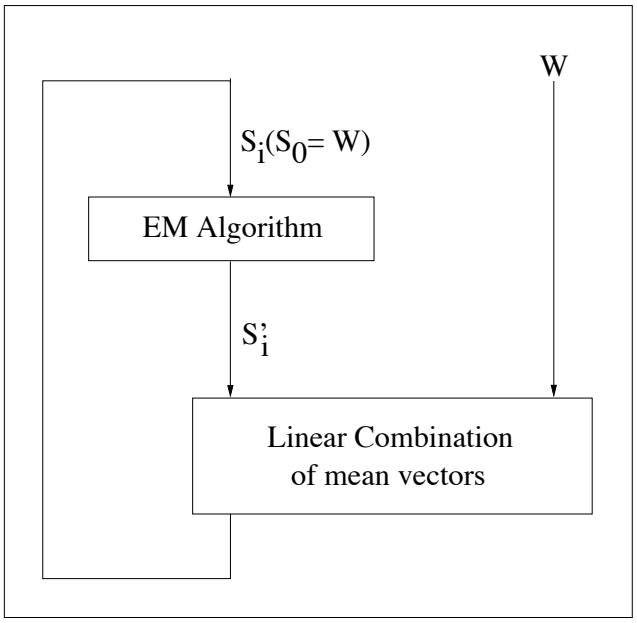

Fig. 2. *MAP procedure used for the training of a speaker model.*

3.4. World Models

For the world models, gaussian mixture models (GMMs) [14] with 128 components and diagonal covariance matrices are trained on speech utterances from the various *world* subsets of the common evaluation protocol using the EM algorithm [3]. The world models are gender and handset dependent.

3.5. Speaker Models

Speaker models are trained using a MAP procedure described in Fig. 2. The corresponding world model, denoted by W, is used as an initialization. Then, one iteration of the EM algorithm is applied, leading to the model S'. Then a new model S_1 is built with the mean vectors being a linear combination of the mean vectors of the two models W and S':

$$\mu_{S_1}^{(i)} = \frac{0.25 \cdot \pi_W^{(i)} \cdot \mu_W^{(i)} + 0.75 \cdot \pi_{S'}^{(i)} \cdot \mu_{S'}^{(i)}}{0.25 \cdot \pi_W^{(i)} + 0.75 \cdot \pi_{S'}^{(i)}}$$

with $\pi_W^{(i)}$ and $\pi_{S'}^{(i)}$ being the weights of the i-th Gaussian component of models W and S' respectively.

Finally, S_1 is used as the initialization for the second iteration of the EM algorithm. 20 iterations are done that way.

3.6. Evaluation

Results of the various systems are measured by a DET curve [10]. For the speaker verification task, the false alarm rate and the miss rate are defined as follows:

$$\mathcal{R}_{FA} = \frac{\text{Number of impostor utterances wrongly accepted}}{\text{Total number of impostor utterances}}$$

$$\mathcal{R}_{MI} = \frac{\text{Number of client utterances wrongly rejected}}{\text{Total number of client utterances}}$$

3.7. Experiments on the Energy

The first set of experiments concerns the integration of the Δ-log-energy and/or the log-energy in the parameter vectors. Fig. 3 shows the results without any normalization technique. The score used is simply the averaged log-likelihood ratio.

The addition of the Δ-log-energy in the feature vectors gives very similar results to a system using only the cepstral and the Δ-cepstral coefficients. Adding the log-energy to the feature vectors clearly degrades the performance. This latter result is surprising because the energy should be used by the GMMs to make a pre-classification between high and low energy frames, thus enabling a better modeling. One possible explanation is the lack of normalization of the

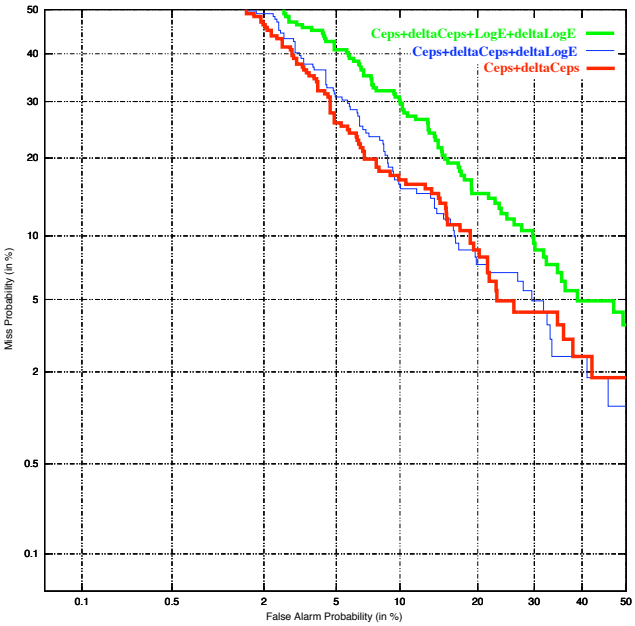

Fig. 3. *Influence of the log-energy and the Δ-log-energy on the performance*.

log-energy. Some kind of normalization of the log-energy should be used to limit the differences between two utterances.

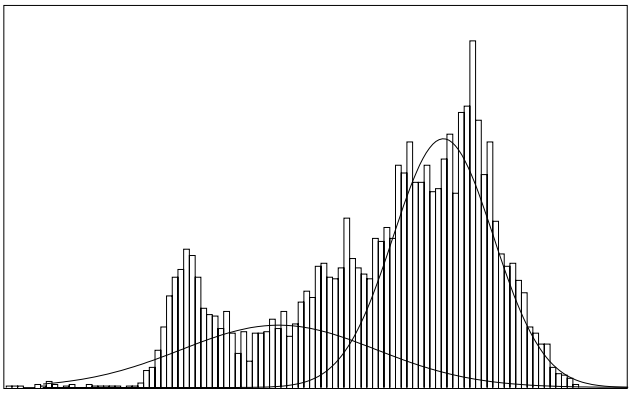

Fig. 4. *Bi-gaussian modeling of the energy distribution*.

3.8. Experiments on Frame Removal

The second set of experiments concerns the study of the influence of frame removal on performance. The method used is based on a bi-gaussian modeling of the energy distribution. First, each frame with a zero energy is automatically discarded. These frames correspond generally to the beginning or the end of a recorded segment, and there are frequently such frames in the Switchboard data. Then, the energy distribution of the remaining frames is calculated, and

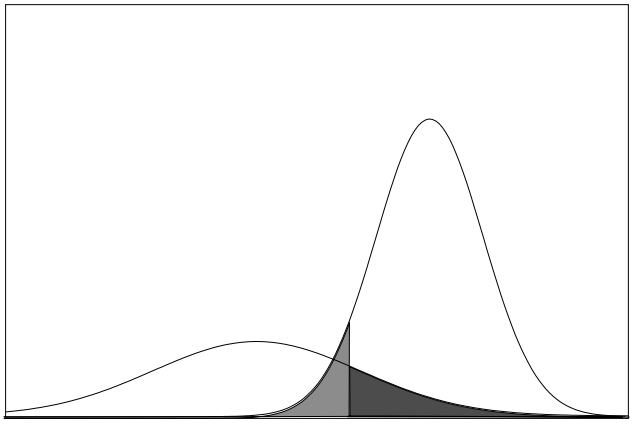

Fig. 5. *Calculation of the threshold on the energy.*

a bi-gaussian model is learned from this distribution using the expectation-maximization (EM) algorithm [3]. Fig. 4 shows an example of energy distribution for a typical test utterance of the NIST speaker recognition evaluation. The bi-gaussian model is also represented on the figure. Once the bi-gaussian model has been estimated, a threshold on the energy is calculated such that the residual surfaces under each gaussian are equal (see Fig. 5). Finally, each frame with an energy below the threshold is discarded.

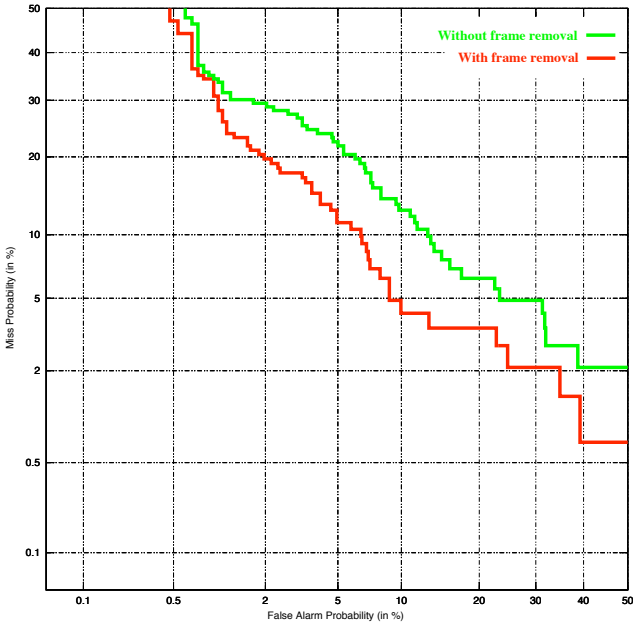

Fig. 6. *Influence of frame removal on the performance*.

Fig. 6 summarizes these experiments and shows results after the application of the h-norm. The application of a frame removal technique improves the performance considerably.

This suggest that the information extracted from low energy frames is not reliable and/or cannot be modeled accurately with GMMs. It therefore helps to simply discard such frames. However, it may be interesting in the future to investigate which, if any, information can be used from those low energy frames.

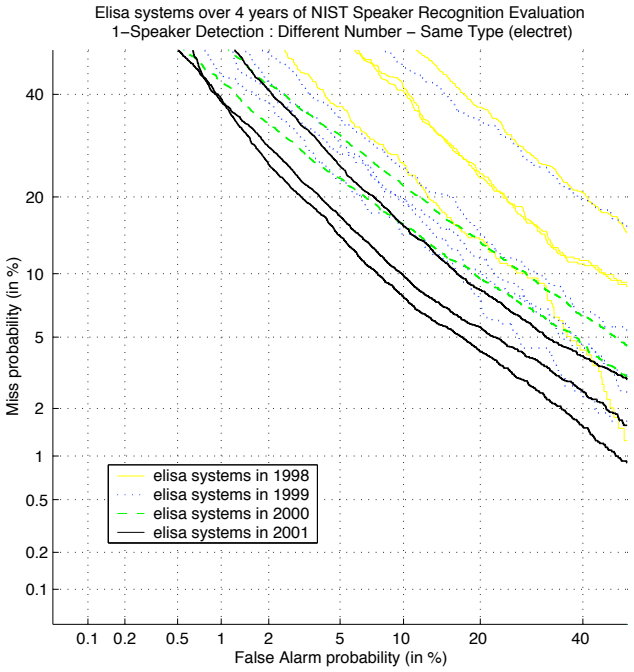

Fig. 7. *ELISA systems over 4 years of NIST speaker recognition evaluation.*

4. RESEARCH DIRECTIONS

Being at the level of the state-of-the-art speaker verification systems allows the Consortium members to investigate on more original research directions. We present the main ones in this section. Some of them are described in details in other papers, other approaches are still currently under investigation.

- Contextual principal components and contextual independent components as an alternative to cepstral analysis for the speech representation module [9];
- Evaluation of the intrinsic quality of a new parameterization using a mutual information criterion;
- Alternative initialization procedures of the world models (vector quantization, mixture of models learned on various random subsets of data, etc.);

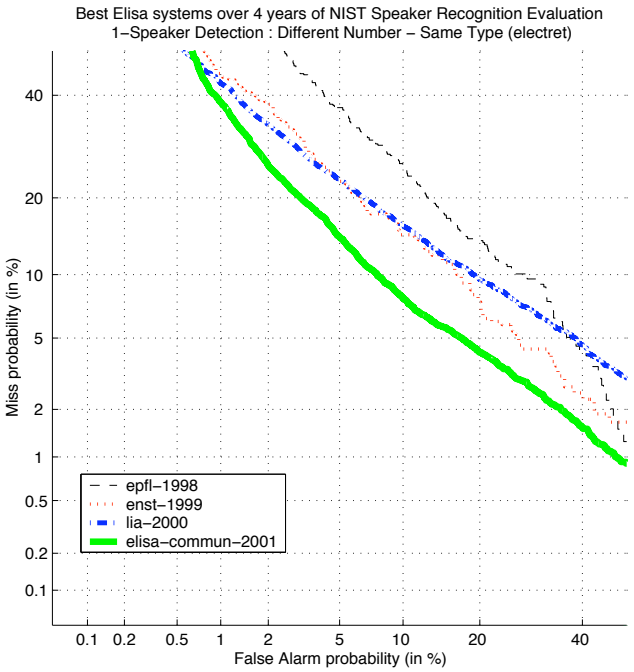

Fig. 8. *Best ELISA systems over 4 years of NIST speaker recognition evaluation.*

- Modeling by a mixture of models, one being trained with an ML procedure, the other being trained with a MAP procedure;
- Alternative MAP procedures to train the speaker models;
- Evaluation of the intrinsic quality of the world and speaker models;
- The use of support vector machines (SVM) as an alternative to log-likelihood-based scoring [7];
- The study of new approaches for score normalization [5];
- Evolutive hidden Markov models for speaker indexing [11].

5. CONCLUSIONS

The ELISA consortium has been created 4 years ago, and a lot of progress has been made since then (see Fig. 7[1] and Fig. 8). For the first time this year, an ELISA platform has been able to provide state-of-the-art performance

[1]The training material and the conditions of the evaluation have changed over the years. Read the evaluation plans provided by NIST on http://www.nist.gov/speech/tests/spk/index.htm for details.

in speaker verification, allowing the members of the consortium to have of a good platform for experimentation in speaker recognition, and leaving time for original research on this topic. One of the main goals of the consortium has been reached thanks to common work between several researchers belonging to various research laboratories. This is an excellent example of collaborative research, and we hope that this year accomplishment will stimulate creative research inside (and outside) the consortium.

6. REFERENCES

[1] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1–3):42–54, January/April/July 2000.

[2] The ELISA Consortium. The ELISA 99 speaker recognition and tracking systems. *Digital Signal Processing*, 10(1–3), January/April/July 2000.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 6(39):1–38, 1977.

[4] Corinne Fredouille, Jean-François Bonastre, and Teva Merlin. AMIRAL: A block-segmental multirecognizer architecture for automatic speaker recognition. *Digital Signal Processing*, 10(1-3):172–197, January/April/July 2000.

[5] Corinne Fredouille, Jean-François Bonastre, and Teva Merlin. Bayesian approach based-decision in speaker verification. In *Proceedings of 2001: A Speaker Odyssey*, June 2001. Crete, Greece.

[6] 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.

[7] Jamal Kharroubi and Gérard Chollet. Text-independent speaker verification using support vector machines. In *ICASSP Student Forum*, May 2001.

[8] Kung-Pu Li and Jack E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *Proceedings of ICASSP 88*, pages 1895–1998, 1988.

[9] Ivan Magrin-Chagnolleau, Geoffrey Durou, and Frédéric Bimbot. Application of time-frequency principal component analysis to text-independent speaker identification. *Accepted for publication in IEEE Transactions on Speech and Audio Processing*.

[10] Alvin 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] Sylvain Meignier, Jean-François Bonastre, and Stéphane Igounet. E-HMM approach for learning and adapting sound models for speaker indexing. In *Proceedings of 2001: A Speaker Odyssey*, June 2001. Crete, Greece.

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

[13] Douglas Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proceedings of EUROSPEECH 97*, pages 963–966, 1997. Rhodes, Greece.

[14] Douglas A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1–2):91–108, August 1995.