SPEAKER DETECTION USING MULTI-SPEAKER AUDIO FILES FOR BOTH ENROLLMENT AND TEST

Jean-François Bonastre, Sylvain Meignier, Teva Merlin

LIA-Avignon – BP1228 – 84911 Avignon Cedex 9 – France

(jean-francois.bonastre, sylvain.meignier, teva.merlin)@lia.univ-avignon.fr

ABSTRACT

This paper focuses on speaker detection using multi-speaker files both for the enrollment phase and for the test phase. This task was introduced during the 2002 NIST speaker recognition evaluation campaign. Enrollment data is composed of three two-speaker files. Test files are also two-speaker records. The system presented here uses a speaker segmentation process based on an HMM conversation model followed by a speaker matching technique to produce one-speaker segments. Speaker detection is then achieved using AMIRAL, LIA's GMM-based speaker verification system. Validation of the proposed strategy is done using extracts from the NIST 2002 results.

1. INTRODUCTION

This paper presents the LIA strategy for the new two speaker task introduced for the NIST 2002 speaker recognition evaluation (2Sp task). The 2Sp task consists in learning a target speaker model using three two-speaker audio files. No information is provided other than that the target speaker is the only speaker presents in each of the three files. Test files are composed of two-speaker records. The Figure 1 synthesizes the 2Sp task.

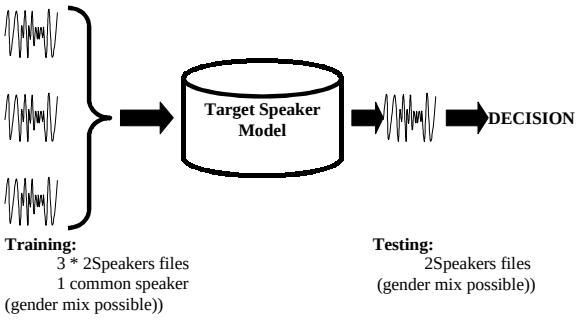

Figure 1: Synopsis of NIST 2speakers task

To deal with this new task, we proposed an approach based on three phases:

- Phase 1 uses the LIA's automatic speaker segmentation system [3]. From each two-speaker file, the system yields two one-speaker subsets.
- Phase 2 concerns only enrolment. A speaker tying/matching process [4] is applied on the six one-speaker subsets issued from the training files of a given speaker. It selects the three subsets relative to the same speaker, which are then used to train the speaker model.
- Phase 3 is a classical one speaker detection process on each of the two one-speaker subsets corresponding to a two-speaker test file. This part uses AMIRAL, the LIA's speaker recognition system [1].

The advantage of the method is that each phase is done using a validated approach in the best conditions. Phase 1 is based on a validated approach for speaker segmentation (witch obtained the best results on Switchboard data during NIST 2002 Campaign) with the associated specific parameterization. Phase two takes into account the specifics of NIST 2Sp task to improve matching performance. Finally, phase three uses a classical speaker verification system, also evaluated using NIST 1998 to 2002 evaluation campaigns.

Section 2 presents the speaker segmentation process, used to split enrolment and test files. Section 3 describes the tying/matching phase. Section 4 is dedicated to the one-speaker verification system. Section 5 presents the experimental conditions and the performance of the proposed approach, in the framework of NIST 2002 evaluation. Finally section 6 concludes and proposes some possible improvements.

2. SPEAKER SEGMENTATION AND SPLITTING

2.1 Speaker segmentation of the audio files

The test and train records get segmented into at most two speakers using LIA's segmentation system [3]. This system models the conversation between the two speakers with a Hidden Markov Model (HMM). Each state of the HMM corresponds to a speaker and the transition model the changes between the two speakers.

The system is initialized by tying a 3s segment of the file to one speaker and remaining to the other speaker. The initial segment is selected as to maximize the likelihood of a Gaussian mixture model (GMM, [6]) learnt on the whole file (through *maximum a posteriori* - MAP – adaptation of a background model (see section 4 for details), in order to insure that the content is speech.

The segmentation is obtained through an iterative process. An iteration starts with the adaptation of the two speaker models to the current segmentation. A new segmentation is then computed using a Viterbi decoding. This process gets repeated as long as the probability of the Viterbi path increased.

The segmentation system needs a specific parameterization optimized for one-file processing (no mismatch). The acoustic parameterization is carried out using the SPRO module developed by the ELISA consortium (20 linear cepstral coefficients and energy, no Δ-cepstral) [2].

2.2 Splitting the audio files

After finding the segments relative to each speaker, the signal is split into two parts according to the segmentation.

Acoustic parameterization (16 linear cepstral coefficients and 16 Δ-cepstral, CMS and variance normalization) is carried out separately on each part of the file (using SPRO).

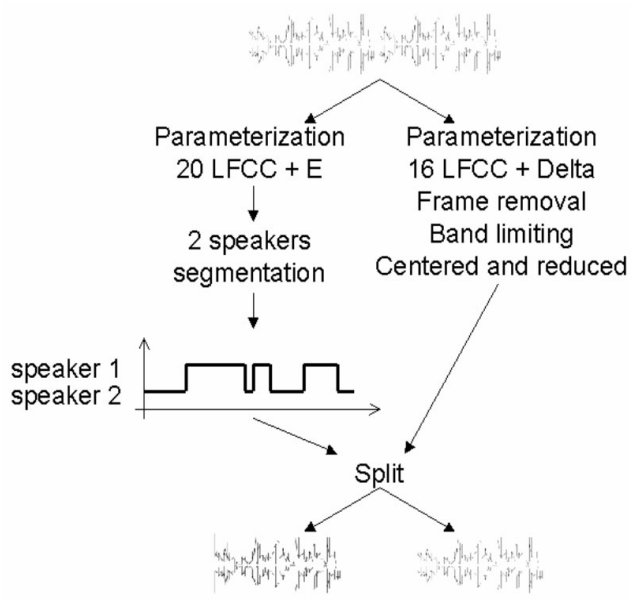

Figure 2: Segmentation and splitting

3. SPEAKER MATCHING

The second step of the proposed method consists in finding the target speaker segments present in each of the three training records. Speaker matching (a.k.a. speaker tying) is a classification problem similar to speaker clustering [5][6].

As proposed in [4], a hierarchical clustering is performed to obtain the three segments. At each stage, the algorithm groups the two closest clusters of segments, according to the log of cross likelihood ratio (CLR) measure [6]. The log of the cross likelihood ratio is given by:

$$l\_clr(x_2, y_1) = \log\left(\frac{l(x_2|Y_1)}{l(x_2|UBM)}\right) + \log\left(\frac{l(y_1|X_2)}{l(y_1|UBM)}\right)$$

where *UBM* is the background model, *l()* the likelihood function, *x* and *y* are 2 files, x_2 and y_1 respectively a segment of *x* and *y*, X_2 and Y_1 the models adapted from *UBM* using respectively x_2 and y_1.

The models used here to compute the CLR are of the same kind as for the speaker detection task which constitutes the final step of the whole process (see section 4) except for the number of components (128 here, 256 for speaker verification task).

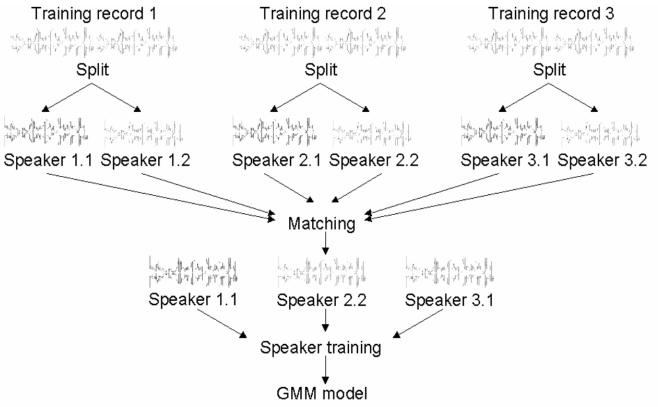

Figure 3: Speaker Matching

Fusion of the segments is severely constrained in this application. The segmentation is considered as accurate and does not get challenged here. So, only two utterances – an utterance is composed of all the segments of a file, labeled as the same speaker – from two different records can be merged. Since we know the target speaker is present in the three training files, the algorithm stops after the second fusion, i.e. the three segments are found. Figure 3 presents the specific matching task realized for this work (up to the speaker training phase described in the next section)

4. SPEAKER VERIFICATION

Thanks to the splitting phase, we only have to deal with a bunch of one-speaker test files (or utterances) to make the final decision. And the matching phase allowed us to identify a subset of utterances representing each target speaker. We now have to build a target model on this

subset and to carry out speaker recognition tests between the model and the two utterances of each test file. If one of the verification test is positive, we decide to accept the test else we reject it. The process is shown figure 4.

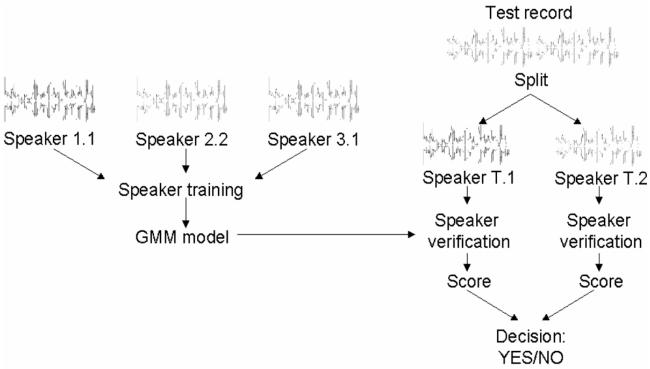

Figure 4: Speaker verification

Both model learning and scoring rely on the use of a universal background model (UBM [6]), which is learnt using a large set of files and EM algorithm respecting Maximum Likelihood criterion. The background model is a GMM with 256 components for this phase (whereas it was 128 for the segmentation and matching phases) and diagonal covariance matrices.

The target models are then adapted from the background model [8]. The adaptation scheme is based on the *maximum a posteriori* method (MAP) [1].

To a score a given test, between a signal utterance and a target speaker, a classical log likelihood ratio is computed for every frame of the utterance using the target model (numerator), the UBM model (denominator).

5. EXPERIMENTS AND RESULTS

5.1 Database

The proposed approach was experimented during NIST 2002 evaluation campaign. The data provided was an extract of SwitchBoard Cellular phase II [http://nist.gov/speech/tests/spk/2002]. Training and test files contained 2-speaker telephonic conversations, including mixed gender conversations.

As seen previously, for each target speaker the training consisted of three conversations with only one speaker, the target, appearing in all three conversations. The three other speakers involved were different, and could be of any gender, with no information available about it. The set of target speaker was composed of 131 males and 178 females.

Test set consisted of 1460 segments with an average duration of 1 minute. As for the training records, the genders within the same record could be unique or mixed.

The world / background model was trained on a different data set composed of records extracted from NIST 2001 development set (SwitchBoard landline corpus).

The evaluation consisted in scoring 22 trials for each test segments, with two of the trials corresponding to the target speaker (the 20 others being impostor speakers). Use of this information was not allowed though.

5.2 segmentation results

To evaluate the segmentation accuracy, table 1 presents results obtained for the NIST 2002 segmentation task. It has to be noted however that the data set was different from the one used for 2Sp. The evaluation method is the NIST official scoring (version 07). It is a frame based error rate protocol.

Missed Speaker Time	0.0%
False alarm Speaker Time	0.0%
Speaker Error Time	7.4%

Table 1: LIA NIST2002 speaker segmentation results for Switchboard corpus

5.3 1-speaker results

In order to evaluate the accuracy of AMIRAL, the one-speaker verification system used for the 2Sp task, we present in figure 5 the results obtained during NIST 2002 evaluation. To help the comparison, results of the same system using landline telephonic data (NIST 2000/2001 data) are also provided.

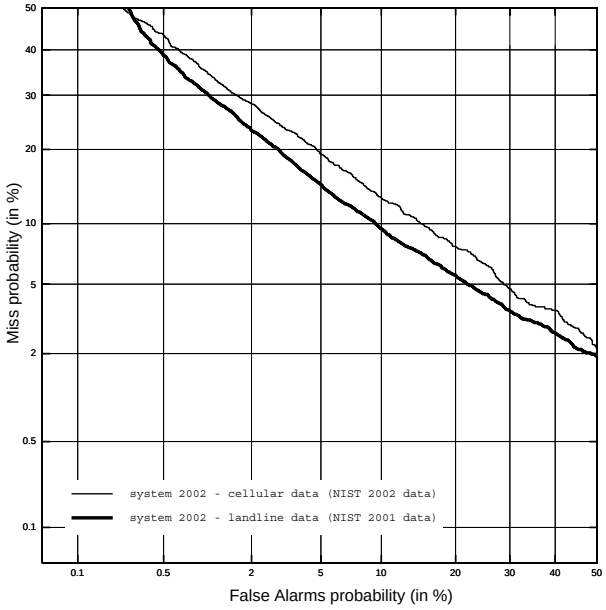

Figure 5: LIA 1-speaker results, NIST 2002 system on NIST 2002 cellular data and NIST 2001 landline data

5.4 Results for the 2Sp task

The results obtained at NIST 2002 evaluation by the 2-speaker system described here are given in figure 6.

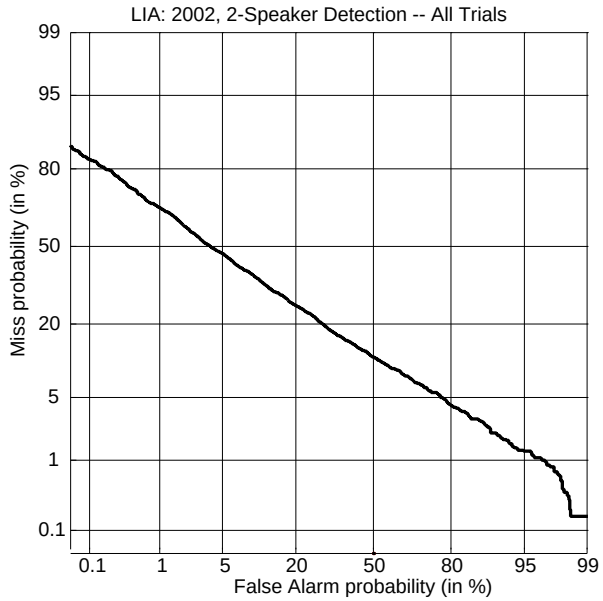

Figure 6: LIA 2-speaker result, NIST 2002 evaluation, all trials.

5.5 Comments

Looking at figures 5 and 6, a large decrease in performance is observed, between the one-speaker system and the two-speaker system (both on cellular data): the equal error rate is approximately twice as high for the latter. A similar loss was observed by all the participants during the last NIST evaluation campaign.

The loss comes for a large part from the different amount of knowledge allowed for the two conditions: for one speaker, the gender and handset are known.

The segmentation process seems accurate enough for the task. As shown in table 1, it achieves a frame error rate of 7.5%; but further analysis of the errors hows that a large part is related to non useful instants of speech (short events or noisy events).

The last source of mistake is the matching process. As a mistake in the matching process leads to a bad target model, its weight in the final score is important.

6. CONCLUSION

In this paper, we propose a complete system for NIST two-speakers task. The target models are learnt using several multi-speaker files and the tests are also done using multi-speaker files.

The proposed solution combines the LIA speaker segmentation, matching, and speaker verification systems.

No specific development was done for the two-speaker task, except for the matching phase, which is a direct extension of our speaker tying system [4]. The system parameters were tuned based on LIA's knowledge in one-speaker detection.

The proposed solution achieves correct results, close to the best NIST 2002 two-speaker system. However, a large loss in performance between one-speaker and two-speaker verification systems is noticed. Further investigations should be pushed concerning the accuracy of the matching process, as the influence of this part of the system is crucial. A better optimization of the system will certainly help reducing the gap between one- and two-speaker results.

6. REFERENCES

[1] Corinne Fredouille, Jean-François Bonastre, and Teva Merlin, "AMIRAL: a block-segmental multi-recognizer architecture for automatic speaker recognition," *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[2] Ivan Magrin-Chagnolleau, Guillaume Gravier, and Raphaël Blouet for the ELISA consortium, "Overview of the 2000-2001 ELISA consortium research activities," in *2001: A Speaker Odyssey*, pp.67–72, Chania, Crete, June 2001.

[3] Sylvain Meignier, Jean-François Bonastre, and Stéphane Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *2001: A Speaker Odyssey*, pp.175-180, Chania, Crete, June 2001.

[4] Sylvain Meignier, Jean-François Bonastre, and Ivan Magrin-Chagnolleau, "Speaker utterances tying among speaker segmented audio documents using hierarchicalclassification: towards speaker indexing of audio databases," in *Proceedings of ICSLP 2002*, Vol. 1, pp 573-576, Denver, Colorado, United States, September 2002.

[5] Douglas A. Reynolds, Elliot Singer, Beth A. Carlson, Gerald C. O'Leary, Jack J. McLaughlin, and Marc A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proceedings of ICSLP 1998*, Sydney, Australia, December 1998.

[6] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, August 1995.

[7] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[8] A. Solomonoff, A. Mielke, M. Schmidt and H. Gish "Clustering speakers by their voice" *in Proceedings of ICASSP 98*, 1998