Step-by-step and integrated approaches in broadcast news speaker diarization

Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-François Bonastre,
Laurent Besacier

HAL Id: hal-01318554
https://hal.science/hal-01318554

Submitted on 24 Mar 2017

Step-by-step and *Integrated* approaches in broadcast news speaker diarization

Sylvain Meignier [a,c] Daniel Moraru [b] Corinne Fredouille [a]
Jean-François Bonastre [a] Laurent Besacier [b]

[a] *LIA/CNRS - University of Avignon - BP1228 - 84911 Avignon Cedex 9 - France*
[b] *CLIPS - IMAG (UJF & CNRS) - BP 53 - 38041 Grenoble Cedex 9 - France*
[c] *LIUM/CNRS, Université du Maine, Avenue Laennec
72085 Le Mans Cedex 9, France*

Abstract

This paper summarizes the collaboration of the LIA and CLIPS laboratories on speaker diarization of broadcast news during the spring NIST Rich Transcription 2003 evaluation campaign (NIST-RT'03S). The speaker diarization task consists in segmenting a conversation into homogeneous segments which are then grouped into speaker classes.

Two approaches are described and compared for speaker diarization. The first one relies on a classical two-step speaker diarization strategy based on detection of speaker turns followed by a clustering process, while the second one uses an integrated strategy where both segment boundaries and speaker tying of the segments are extracted simultaneously and challenged during the whole process. These two methods are used to investigate various strategies for the fusion of diarization results.

Furthermore, segmentation into acoustic macro-classes is proposed and evaluated as a prior step to speaker diarization. The objective is to take advantage of the a-priori acoustic information in the diarization process. Along with enriching the resulting segmentation with information about speaker gender, channel quality or background sound, this approach brings gains in speaker diarization performance thanks to the diversity of acoustic conditions found in broadcast news.

The last part of this paper describes some ongoing works carried out by the CLIPS and LIA laboratories and presents some results obtained since 2002 on speaker diarization for various corpora.

Key words: Speaker Indexing, Speaker Segmentation and Clustering, Speaker Diarization, E-HMM, Integrated approach, Step-by-step approach

Email addresses: `sylvain.meignier@univ-lemans.fr` (Sylvain Meignier),

1 Introduction

The design of efficient indexing algorithms to facilitate the retrieval of relevant information is vital to provide easy access to multimedia documents. Until recently, indexing audio-specific documents such as radio broadcast news or the audio channel of video materials mostly consisted in running automatic speech recognizers (ASR) on the audio channel in order to extract syntactic or higher level information. Text-based information retrieval approaches were then applied to the transcription issued from speech recognition. The transcription task alone represented one of the main challenges of speech processing during the past decade (see the DARPA workshop proceedings at [1]) and no specific effort was dedicated to other information embedded in the audio channel. Progress made in broadcast news transcription [2,3] shifts the focus to a new task, denoted "Rich Transcription" [4], where syntactic information is only one element among various types of information. At the first level, acoustic-based information like speaker turns, the number of speakers, speaker gender, speaker identity, other sounds (music, laughs) as well as speech bandwidth or characteristics (studio quality or telephone speech, clean speech or speech over music) can be extracted and added to syntactic information. At the second level, information directly linked to the spontaneous nature of speech, like disfluencies (hesitations, repetitions, etc.) or emotion is also relevant for rich transcription. On a higher level, linguistic or pragmatic information such as named entity or topic extraction for instance is particularly interesting for seamless navigation or multimedia information retrieval. Finally, some types of information extraction relevant to document structure do not fall exactly into one category; for example, the detection of sentence boundaries can be based on acoustic cues but also on linguistic ones.

This paper concerns information extraction on the first level described above. It is mainly dedicated to the detection of speaker information, such as speaker turns, speaker gender, and speaker identity. These speaker-related tasks correspond to speaker segmentation and clustering, also denoted speaker diarization in the NIST-RT evaluation terminology.

The speaker diarization task consists in segmenting a conversation involving multiple speakers into homogeneous parts which contain the voice of only one speaker, and grouping together all the segments that correspond to the same speaker. The first part of the process is also called speaker change detection while the second one is known as the clustering process. Generally, no prior information is available regarding the number of speakers involved or their

daniel.moraru@imag.fr (Daniel Moraru),
corinne.fredouille@lia.univ-avignon.fr (Corinne Fredouille),
jean-francois.bonastre@lia.univ-avignon.fr (Jean-François Bonastre),
laurent.besacier@imag.fr (Laurent Besacier).

identities. Estimating the number of speakers is one of the main difficulties for the speaker diarization task. To summarize, this task consists in:

- finding the speaker turns,
- grouping the speaker-homogeneous segments into clusters,
- estimating the number of speakers involved in the document.

Classical approaches for speaker diarization [5–8] deal with these three points successively: first finding the speaker turns using by example the symmetric Kullback Leibler (KL2), the Generalized Likelihood Ratio (GLR), or the Bayesian Information Criterion (BIC) distance approaches, then grouping the segments during a hierarchical clustering phase, and finally estimating the number of speakers a posteriori. If this strategy presents some advantages like dealing with quite long and pure segments for the clustering, it also has some drawbacks. For example, knowledge issued from the clustering (like speaker-voice models) could be very useful for segment boundary estimate as well as to facilitate the detection of other speakers. Contrasting with this *step-by-step* strategy, an *integrated* approach in which the three steps involved in speaker diarization are performed simultaneously, allows to use all the information currently available for each of the subtasks [9,10]. The main disadvantage of the *integrated* approach lies in the need to learn robust speaker models using very short segments (rather than a cluster of segments as in classical approaches), even though the speaker models get refined along the process. Mixed strategies are also proposed [6,11,12], where classical *step-by-step* segmentation and clustering are first applied and then refined using a "re-segmentation" process during which the segment boundaries, the segment clustering and sometimes the number of speakers are challenged jointly.

In addition to the intrinsic speaker diarization subtasks presented above (denoted p1 in the list below), various problems need to be solved in order to segment an audio document into speakers, depending on the environment or the nature of the document:

- to identify the speaker turns and the speaker clusters, and to estimate the number of speakers involved in the document, without any a-priori information (p1);
- to be able to process speech documents as well as documents containing music, silence, and other sounds (p2);
- to be able to process spontaneous speech with overlapping voices of speakers, disfluencies, etc. (p3).

The NIST'02 speaker recognition evaluation provided an overview of the performance that can be obtained for:

- conversational telephone speech, involving two speakers and a single acoustic class of signals ;

- broadcast news data which often includes various qualities or types of signal (such as studio/telephone speech, music, speech over music, etc.) ;
- meeting room data in which speech is more spontaneous than in the previous cases, and presents several distortions due to distant microphones (*e.g.* table microphone) and noisy environment.

Table 1 shows the various classes of problems encountered in each situation (p1, p2, and p3). The increasing difficulty of the tasks is obviously due to their novelty (the last two tasks were introduced for the 2002 evaluation campaign) but also and mainly to the accumulation of problems described in the previous paragraph.

Since 2001, two members of the ELISA Consortium, CLIPS and LIA, have been collaborating in order to participate in the yearly evaluation campaigns for the task of speaker segmentation/diarization: NIST'01 [13] (LIA only), NIST'02 [14], NIST-RT'03S [4], and NIST-RT'04S [15]. Since speaker diarization may also be useful for indexing and segmenting videos, CLIPS has also participated in experiments in the last three TREC VIDEO evaluations [16] since 2002 [17,18].

The ELISA Consortium was originally created by ENST, EPFL, IDIAP, IRISA and LIA in 1998 with the aim of promoting scientific exchange between members, developing a common state-of-the-art speaker verification system and participating in the yearly NIST speaker recognition evaluation campaigns. With the years, the composition of the Consortium has changed and today CLIPS, DDL, ENST, IRISA, LIA, LIUM and the Friburg University are members. Since 1998, the members of the Consortium have participated in the NIST evaluation campaigns in speaker verification; a comparative study of the various systems presented in the 1999, 2000 and 2001 campaigns can be found in [19,20].

This paper presents an overview of this long-term collaboration by investigating two main issues. Firstly, the relative advantages of the classical *step-by-step* approach as well as of a more original *integrated* strategy are discussed (this part of the work can be linked to the "p1" point mentioned above: the intrinsic tasks of speaker diarization). Several fusion strategies that use the advantages of both approaches are also proposed. The second issue addressed in this paper concerns the nature of the audio documents to be segmented (issue denoted as "p2"). This part of the work is more precisely dedicated to speaker diarization of broadcast news data. The interest of applying an acoustic macro-class segmentation process before speaker segmentation (in order to divide the audio file into bandwidth- or gender-homogeneous parts) are discussed.

This paper is organized as follows: section 2 is devoted to the description of systems. The acoustic macro-class segmentation process and the two speaker

diarization techniques are described successively. Section 3 focuses on the fusion of the two approaches. Performance of the various systems are presented and discussed in section 4. All the experimental protocols and data are taken from the NIST-RT'03S development and evaluation corpora (except for some results on meeting data reported in section 5, issued from the NIST-RT'04S meeting data evaluation [15,21]). Section 5 presents ongoing work on meeting data and integration of a-priori knowledge into a speaker diarization system. Finally, concluding remarks are made in section 6.

2   Speaker diarization approaches

Two different speaker diarization systems are proposed in this paper and described in the next sections. They were developed individually by the CLIPS and LIA laboratories in the framework of the ELISA consortium [22,12]. The CLIPS system relies on a classical *step-by-step* strategy. It involves a distance based detector strategy [23] followed by a hierarchical clustering. This approach will be denoted as *step-by-step* strategy in the rest of this paper. The second system developed by the LIA follows an integrated strategy. It is based on a HMM and will be denoted as *integrated* strategy in this paper.

As illustrated in figure 1, both of them use an acoustic macro-class segmentation as a preliminary phase. During this acoustic segmentation, the signal is first divided into four acoustic classes according to different conditions based on gender and wide/narrow band detection. Then, the (CLIPS and LIA) diarization systems are individually applied on each isolated acoustic class. Finally, the four resulting segmentation outputs are merged and consolidated through a re-segmentation phase. The separate application of the speaker diarization systems on each acoustic class assumes that a particular speaker is associated with one of them only. Nevertheless, the re-segmentation process allows to question the relationship between a speaker and a unique acoustic class.

Both diarization approaches and acoustic segmentation were developed independently before to investigate different strategies for combining the systems. The settings of each of them, as acoustic features or learning methods, differ and came from experiments conducted over a common development corpus (see section 4.1).
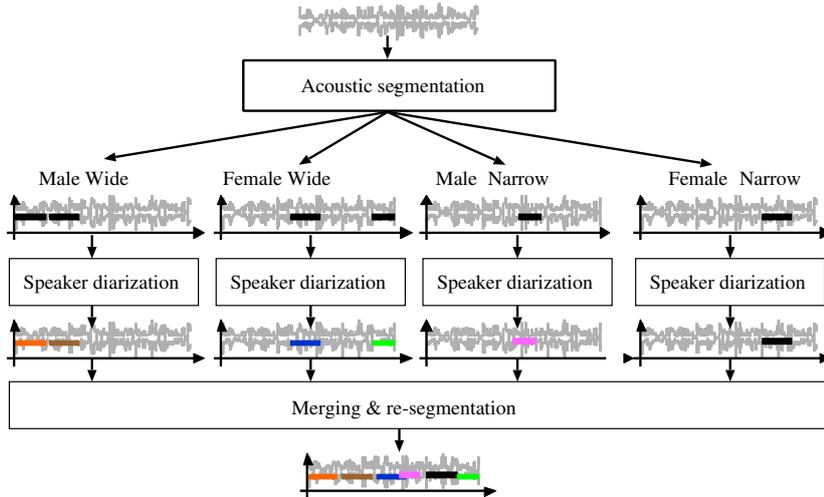
Fig. 1. *Overview of the speaker diarization strategy*

2.1 Acoustic macro-class segmentation

Segmenting an audio signal into acoustic classes was mainly introduced to assist automatic speech recognition (ASR) systems within the special context of broadcast news transcription [24–26]. Indeed, one of the first objectives of acoustic segmentation was to provide ASR systems with an acoustic event classification to discard non-speech signal (silence, music, advertisements) and to adapt ASR acoustic models to some particular acoustic environments, like speech over music, telephone speech or speaker gender. Many papers were dedicated to this particular issue and to the evaluation of acoustic segmentation in the context of the ASR task. However, acoustic segmentation may be useful for other tasks linked to broadcast news corpora, although this is rarely discussed in the literature. In this sense, one of the aims of this work is to investigate the impact of acoustic segmentation when it is applied as prior segmentation for speaker diarization.

Speech/non-speech detection is useful for the speaker diarization task in order to avoid music and silence portions being automatically labeled as new speakers. This is particularly true in the context of the NIST-RT evaluation in which both miss and false alarm speech errors are taken into account for the speaker diarization scoring.

Moreover, an acoustic segmentation system can be designed to provide a finer classification. For example, gender and frequency band detections may introduce a-priori knowledge in the diarization process. In this paper, the prior acoustic segmentation is done at three different levels:

- Speech / non-speech.
- Clean speech / Speech over music / Telephone speech (narrow band).

- Male / Female speech.

2.1.1 Hierarchical approach

The system relies on a hierarchical segmentation performed in three successive steps using Viterbi decoding as illustrated in figure 2:

- During the first step, a speech / non-speech segmentation is performed using two models. The first model, *MixS*, represents all the speech conditions while the second one, *NS*, represents the non-speech conditions. Basically, the segmentation process relies on a frame-by-frame best model search. A set of morphological rules is then applied to aggregate frames and to label segments. These rules mainly aim at constraining the duration of segments, by fixing for instance minimum lengths for both speech and non-speech segments. This strategy was preferred to a Viterbi decoding, which tends to misclassify non-speech segments.
- During the second step, a segmentation based on 3 classes, clean speech (S model), speech over music (SM model) and telephone speech (T model), is performed only on the speech segments detected during the previous segmentation step. All the models involved during this step are gender-independent. The segmentation process is a Viterbi decoding applied on an ergodic HMM, composed of three states (S, T, and SM models). The transition probabilities of this ergodic HMM are learnt on the 1996 HUB 4 broadcast news corpus.
- The last step is gender detection. According to the label assigned during the previous step, each segment will be identified as female or male speech by the use of models dependent on both gender and acoustic classes. *GT-Fe* and *GT-Ma* models represent female and male telephone speech respectively, *GS-Fe* and *GS-Ma* represent female and male clean speech, while *GSM-Fe* and *GSM-Ma* represent female and male speech over music. Two additional models, *GDS-Fe* and *GDS-Ma*, representing female and male speech recorded under degraded conditions are also used to refine the final segmentation. The segmentation process described in the previous step is applied here again.

2.1.2 System specifications

The signal is characterized by 39 acoustic features computed every 10 ms on 25 ms Hamming-windowed frames: 12 Mel Frequency Cepstral Coefficients (MFCC) augmented by the normalized log-energy, followed by the delta and delta-delta coefficients. The choice of parameters was mainly guided by the literature [24].
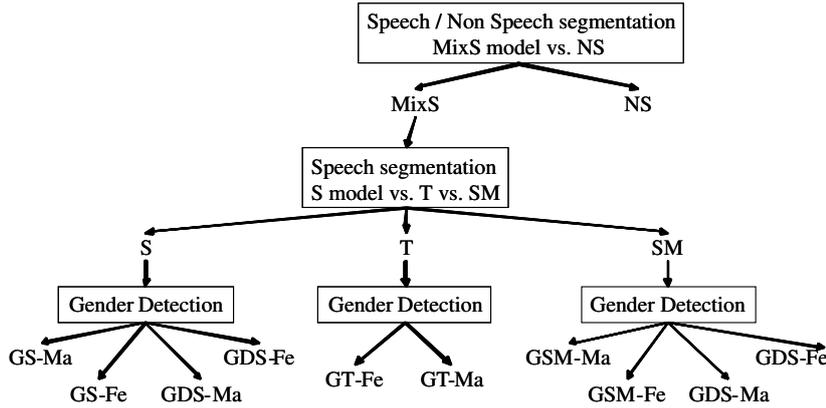
Fig. 2. *Hierarchical acoustic segmentation*

All the models mentioned in the previous section are diagonal GMMs, trained on the 1996 HUB 4 broadcast news corpus. The *NS* and *MixS* models are characterized by 1 and 512 Gaussian components respectively, while the other models are characterized by 1024 Gaussian components. All these parameters have been chosen empirically following a set of experiments not reported here.

2.2 Step-by-step speaker diarization

The CLIPS system is a state-of-the-art system based on speaker change detection followed by a hierarchical clustering. The number of speakers involved in the conversation is automatically estimated. The system uses the same acoustic macro-class segmentation as the LIA system. The CLIPS diarization is applied individually on every acoustic class as explained in section 2 and the results are merged at the end. The next subsections will provide a detailed description of every module of the system.

2.2.1 Step One: Speaker Change Detection

The goal of speaker change detection is to cut the audio recording into segments containing only the speech of one single speaker. The purpose of speaker change detection is to find some audio signal discontinuities that will help us distinguish between two consecutive speakers. Those segments will be used as input data for the clustering module. A distance based approach [23,8]) is used. The Generalized Likelihood Ratio (GLR) is used in our case. Given two acoustic sequences X and Y we ask whether they were produced by the same gaussian model (the same speaker) M_{XY} or by two different models (two different speakers) M_X and M_Y. This question can be answered using the following GLR ratio, where:

$$R_{GLR}(X;Y) = logL(X|M_X) + logL(Y|M_Y) - logL(XY|M_{XY}) \qquad (1)$$

A high value of R_{GLR} means that the "two-model hypothesis" is more likely than the "one-model hypothesis". The numerator of R_{GLR} is the log-likelihood of the "two-model hypothesis" and the denominator is the log-likelihood of the "one-model hypothesis". A GLR curve is extracted from 1.75-second adjacent windows that move along the audio signal. The window size must be small enough to contain only one speaker and large enough to obtain a reliable model. The two windows advance frame by frame. Mono-Gaussian models with diagonal covariance matrices are used to build the GLR curve. The maximum peaks of the curve are the most likely speaker change points. A threshold is then applied on the GLR curve to find speaker changes. The threshold is tuned so that over-segmentation (more speaker changes detected) is provided, as we prefer to detect more segments (which can be further merged by the clustering process) rather than miss speaker changes (which will never be recovered later). The threshold is computed using the mean value of the current curve. Thus, it does not create any adaptation problems from one file to another.

Another system was presented at the NIST'02 speaker recognition evaluation with a-priori segmentation using fixed length segments (0.75 second). It gave approximately the same performance while being 3 times slower due to the uniform segmentation that leads to far more segments as input of the clustering module.

2.2.2 Step Two: Clustering

Now that we have detected the speaker changes, the segments obtained must be grouped (clustered) by speaker. The CLIPS clustering uses a hierarchical bottom-up algorithm. A clustering algorithm generally relies on two important things: the distance between classes and the stop criterion. The distance used is the GLR distance and the stop criterion is the estimated number of speakers. The GLR distance is the GLR ratio (already mentioned in the previous section) computed between classes rather than consecutive windows. Another difference is that the models used are no longer mono-gaussian as in speaker change detection but Gaussian Mixture Models (GMMs).

First, a diagonal 32 GMM background model is trained on the entire file using a classical EM algorithm. We need a background model to compensate for the lack of data for each speaker. The advantage of using a background model trained on the current file is that it is always suited for the current task. A more complex background model (e.g 512 GMM diagonal) trained on external data could perform better but makes the speaker diarization system to be data dependent (the system would work only on the type of data used to

train the background model). The size of the model is a good compromise between complexity and performance: beyond 32 gaussian component we only gain about 0.5% absolute segmentation error but we increase the execution time.

Segment models are then trained using a linear MAP adaptation of the background model (means only). GLR distances are then computed between models and the closest segments are merged at each step of the algorithm until N segments are left (corresponding to N speakers detected in the conversation).

The number of speakers N is estimated as described in the next section. The clustering is done individually on each acoustic macro-class (namely male/wide, female/wide, male/narrow and female/narrow) and the results are merged in the end.

2.2.3 Step Three: Estimating the number of speakers

The algorithm that estimates the number of speakers is based on penalized Bayesian Information Criterion (BIC) [27].

At first, the number of speakers is limited to between 1 and 25. The upper limit usually depends on the recording size.

We select the number of speakers (N_{sp}) that maximizes:

$$BIC(M) = logL(X|M) - \lambda\frac{m}{2}N_{sp}logN_X \qquad (2)$$

where M is the model composed of the N_{sp} speaker models, N_X is the total number of speech frames involved, m is a parameter that depends on the complexity of the speaker models and λ is a tuning parameter empirically set at 0.6. In our case (32 diagonal GMM) m is equal to 64 (2 times 32) times the number of acoustic features. The first term is the overall log-likelihood of the data. The second term is used to penalize the complexity of the model. We need the second term because the log-likelihood of the data increases with the number of models (speakers) involved in the calculation of $L(X|M)$.

Let X_i and M_i be the data and the model of speaker i respectively. The model is obtained by MAP adaptation of the background model over the speaker data as in the previous section. If we make the hypothesis that data X_i depends only on the speaker model M_i then we can prove that the overall log-likelihood of the data becomes:

$$L(X|M) = \prod_{i=1}^{N_{sp}} L(X_i|M_i) \qquad (3)$$

Results concerning the estimation of the number of speakers will be presented in section 4.

2.2.4   System specifications

The signal is characterized by 16 Mel frequency cepstral features (MFCC) computed every 10ms on 20ms windows using 56 filter banks. Then we add the energy parameter. The choice of the number of filters is due to the fact that we work on wide-band data (broadcast news). No frame removal nor coefficient normalization is applied. The parameterization is the same for all system modules of this step-by-step diarization system, but is different than that of the "integrated" speaker diarization system and the acoustic segmentation, which were all developped separately in different places.

2.3   Integrated speaker diarization

The LIA system is based on an evolutive Hidden Markov Modeling (E-HMM) of the conversation [28,9,22,12]. The HMM is ergodic; all speaker changes are potentially available. Each state of the HMM characterizes a speaker and the transitions model the changes between speakers (Figure 3). In this iterative approach, both the segmentation and the speaker models are used at each step and are re-evaluated in the next step. During the diarization process, the speakers are detected and added one by one at each iteration. This is the reason why we have named this diarization method *integrated* approach.
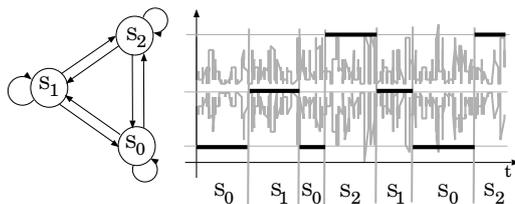
Fig. 3. *Integrated approach : evolutive HMM modeling of the conversation and segmentation. Example given for 3 speakers (S_0, S_1, S_2).*

The speaker diarization system relies on the acoustic macro-class segmentation described in section 2.1. It is applied separately on each of the acoustic classes detected (*e.g.* male/wide, female/wide, male/narrow and female/narrow). Finally, the separate speaker diarizations are merged followed by a re-segmentation process, described in section 2.4.

11

During the diarization, the HMM is generated using an iterative process which detects and adds a new state (*i.e.* a new speaker) at each iteration. The speaker detection process is performed in four steps (figure 4). An example for a two speaker show is given in the figure 5:
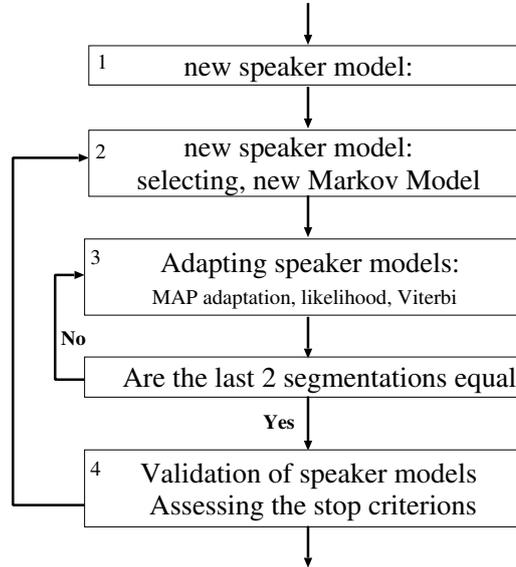
Fig. 4. *Integrated approach: different steps of the process*

Step 1-Initialization
A first speaker model S_0 is trained on the whole test utterance. The segmentation is modeled by a one-state HMM and the whole signal is assigned to speaker S_0. S_0 represents the speaker set of the test record.

Step 2-Adding a new speaker
A new speaker is extracted from the segments currently labeled S_0 representing the speakers that are not detected yet. The new speaker model is trained using the 3-second region of S_0 that maximize the likelihood ratio between model S_0 and a Universal Background Model (UBM [29], see section 2.3.2). The length of the initial region must be sufficient to initialize a robust speaker model while containing one speaker only. This strategy selects the closest data to speaker model S_0. The 3-second length is chosen empirically.

A corresponding state, labeled S_x (x is the number of iterations), is added to the previous HMM. The transition probabilities are updated according to a set of rules (more details are given in section 2.3.2). Finally, the selected 3 seconds of test are moved from label S_0 to label S_x in the segmentation hypothesis.

Various selection strategies have been tested, involving either the speaker or UBM models. The selection method described here produces the best accuracy

in terms of purity of the segments and of speaker diarization error.

Step 3-Adapting speaker models
This phase allows to detect the segments belonging to new speaker S_x and to question the other speakers. First, all the speaker models are adapted according to the current segmentation. Then, Viterbi decoding produces a new segmentation. The adaptation and decoding tasks are performed while the segmentation differs between two successive adaptation/decoding phases. Two segmentations are different when at least a feature is labeled by a different speaker.

Step 4- Speaker model validation and assessment of the stop criterion
The likelihood of the previous solution and the likelihood of the current solution are computed using the current HMM model (for example, the solution with two speakers detected and the current solution with three speakers detected). The stop criterion is reached when no gain in terms of likelihood is observed or when no more speech is left to initialize a new speaker.

During the development, experiments shows that two heuristics help to minimize the speaker diarization error:

- The first one removes the current speaker if the length of its segments is less than 4 seconds. Moreover, the 3-second region used for it initialization is not employed again in the step 2 and the process continues with the segmentation of the previous iteration.
- The second one discards the previous speakers from the segmentation if the length of their segments is lower than the current one. This rule, which forces the detection of the longest speaker first, is closely related to the evaluation metric used in NIST campaigns where it is more important to find the longest speaker segments than the shortest ones.

2.3.2 System specifications

The system specifications are set empirically on a development corpus (see section 4.1). The next paragraphs give some details on the parameterization of the signal, the speaker model adaptation and the HMM.

Parameterization
The signal is characterized by 20th order Mel cepstral features (MFCC) computed at a 10ms frame rate using a 20ms window and the normalized energy. No coefficient normalization is applied; indeed the cepstral mean subtraction (CMS) or the sliding CMS decreases the diarization accuracy.

Speaker models
Speaker models and adaptation techniques used in the E-HMM are similar to
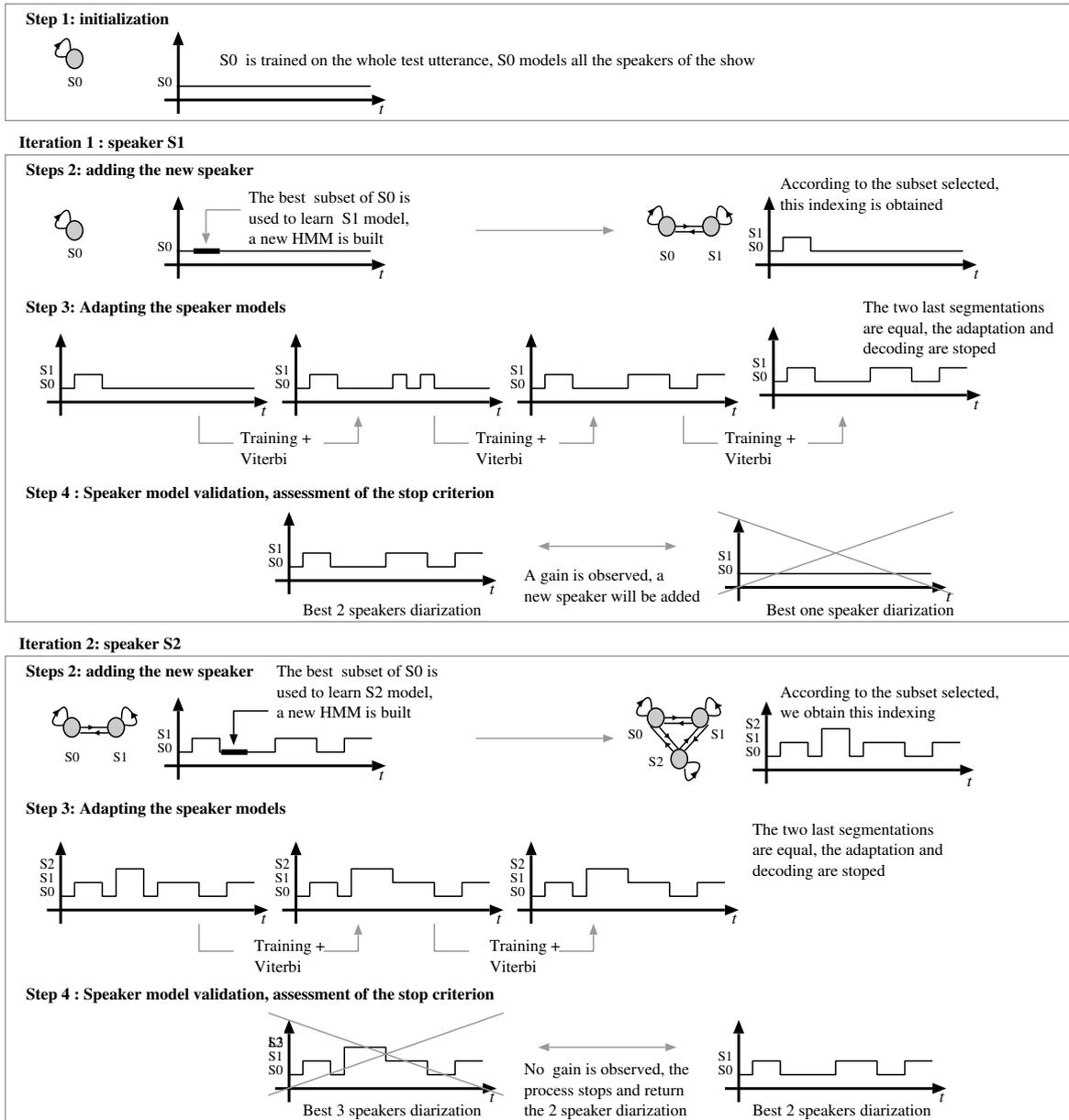
Fig. 5. *Integrated approach: diarization example for a two speaker show*

those generally used for automatic speaker recognition. Speaker models are based on Gaussian Mixture Models (GMM) derived from a UBM. Means only are adapted by a MAP technique. The GMMs are composed of 128 Gaussian components with diagonal covariance matrices.

The UBM is trained with a classical EM algorithm based on the ML principle on a subset of 1996 HUB 4 broadcast news corpus. The UBM learning set is composed of both male and female data and both wide- and narrow-band data. Variance flooring is applied during the training so that variance for each Gaussian is no less than $0.5\times$ the variance of the corresponding UBM Gaus-

sian. The sliding Cepstral Mean Subtraction (CMS) is applied on each training data set before learning ; the sliding window is 3 seconds long. The CMS is performed to remove the influence of the various channels (due to the high number of speakers and records in the UBM corpus). Moreover, preliminary experiments had shown an improvement of the speaker diarization accuracy when the UBM features were normalized.

The adaptation scheme is based on a variant of MAP developed by the LIA [9]. The relative weights of the UBM and the estimate data result from a combination of the UBM and estimated speaker gaussian weights (respectively w_i^{UBM}, w_i^E for the gaussian i) and a-priori weights (respectively $\alpha, 1 - \alpha$). The mean i of the speaker model is obtained by:

$$\mu_i = \frac{\alpha \; w_i^{UBM}}{\alpha \; w_i^{UBM} + (1 - \alpha) \; w_i^E} \; \mu_i^{UBM} + \frac{(1 - \alpha) \; w_i^E}{\alpha \; w_i^{UBM} + (1 - \alpha) \; w_i^E} \; \mu_i^E \qquad (4)$$

Experimentally, α is fixed to 0.2 for the UBM. This setting corresponds to the value that minimizes the speaker diarization error over the development corpus.

HMM
The HMM emission probabilities - for each 0.3 second of the input stream and each HMM state - are estimated by computing the mean of log likelihood. A 0.3-second score rate (the systems are generally based on a frame score rate) allows to smooth out local speaker changes and modifies the intrinsic exponential duration law of the states. Taken a decision every 0.3-second could be also seen as an uniform segmentation.

The HMM transition probabilities are fixed according to the following rules :

- Each transition probability, $a_{i,i}$ (from state S_i to state S_i) is equal to an *a-priori* value η.
- Each transition probability, $a_{i,j}$ (from state S_i to state S_j) is equal to:

$$a_{i,j} = \frac{(1 - \eta)}{(n - 1)} \qquad (5)$$

with $i \neq j$ and n is the number of states (*i.e.* speakers).

In this paper, the η value is set at 0.6. This setting corresponds to the value that minimizes the speaker diarization error over the development corpus.

15

The use of a re-segmentation phase at the end of a clustering process was earlier proposed, for example in [30,31,11,32]. The two main methods are based on GMM/HMM models and make decisions at the frame level:

- thanks to Viterbi decoding [30,31];
- or over scores computed in a sliding window [11,32].

The process can be run iteratively but [11] has shown that it degrades the performance.

The ELISA re-segmentation stage is also based on a Viterbi decoding (similar to the "Adapting speaker models step", described in 2.3.1). Firstly, the four gender- and channel-dependent segmentations are merged by simply pooling the segmentations (there is no overlap between sub-segmentations). Secondly, the speaker-model adaptation and Viterbi decoding are performed iteratively. At the end of each iteration, the speakers with less than 4 seconds of signal are removed.

During the re-segmentation process, the parameters are similar to those used for the E-HMM clustering process, except for the model training method. In this case, the classical mean-only MAP adaptation is performed to obtain speaker models [33,29] instead of the variant MAP technique proposed by the LIA and described in section 2.3.2. The adaptation rate of the means is controlled by the relevant factor [29] which is experimentally set at 16. Moreover a tiny gain is obtained over the development corpus when the HMM emission probability score rate is reduced from 0.3 second to 0.2 second since this reduction helps to refine the boundaries of the output segmentation.

3   Fusion of systems

Since the NIST 2002 evaluation, CLIPS and LIA have investigated different strategies for combining the systems. In this paper, only strategies for broadcast news data are described[1] . Basically, the aim of these strategies is to benefit from the advantages of both speaker diarization approaches, described in previous sections. Two kinds of strategy are proposed: firstly an hybridization strategy and secondly merging various diarizations. The latter is a new way of combining results coming from multiple and unlimited diarization systems.

_____

[1] The reader is invited to look at [22] for telephone strategy.

3.1 Hybridization strategy ("piped" system)

The purpose of this hybridization strategy is to use the results of one system to initialize a second one. In this paper, the speakers detected by the *step-by-step* system (number of speakers and associated audio segments) are inserted in the re-segmentation module of the *integrated* system (the models are trained using the information provided by the clustering phase) as illustrated in figure 6. This solution associates the advantages of longer and (quite) pure segments, provided by the *step-by-step* approach, with the HMM modeling and decoding power of the *integrated* strategy.
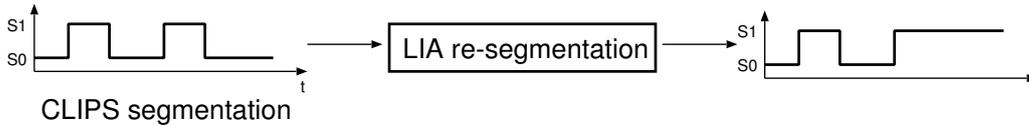

Fig. 6. *Example of a piped system*

3.2 Merging strategy ("fusion" system)

The aim of the "fusion" system consists of using the diarizations issued from as many experts as possible. For example, in this paper the total number of experts is four (see figure 7): the *step-by-step* system, the *integrated* system, a variant of the *integrated* system, and the "piped" system (seen before). The merging strategy relies on a frame-based decision which consists of grouping the labels proposed by each of the systems at the frame level. An example (for four systems denoted A, B, C and D) is illustrated below:

- Frame i: System A gives the speaker label A_1, System B gives B_4, System C gives C_1 and System D gives D_1. $A_1B_4C_1D_1$ is then the merged label.
- Frame i+1 : Systme A gives A_2, System B gives B_4, System C gives C_1 and System D gives D_1. $A_2B_4C_1B_1$ is then the merged label.

This label merging method generates (before re-segmentation) a large set of potential speakers. The re-segmentation module of the *integrated* system can be applied on the merged diarization. Between each adaptation / decoding phase, the potential speakers for whom total time is shorter than 3 seconds are deleted. Indeed, 3 seconds of signal correspond to the minimal length needed to learn a speaker model. The data of these deleted speakers will further be dispatched between the remaining speakers during the next adaptation / decoding phase.

17
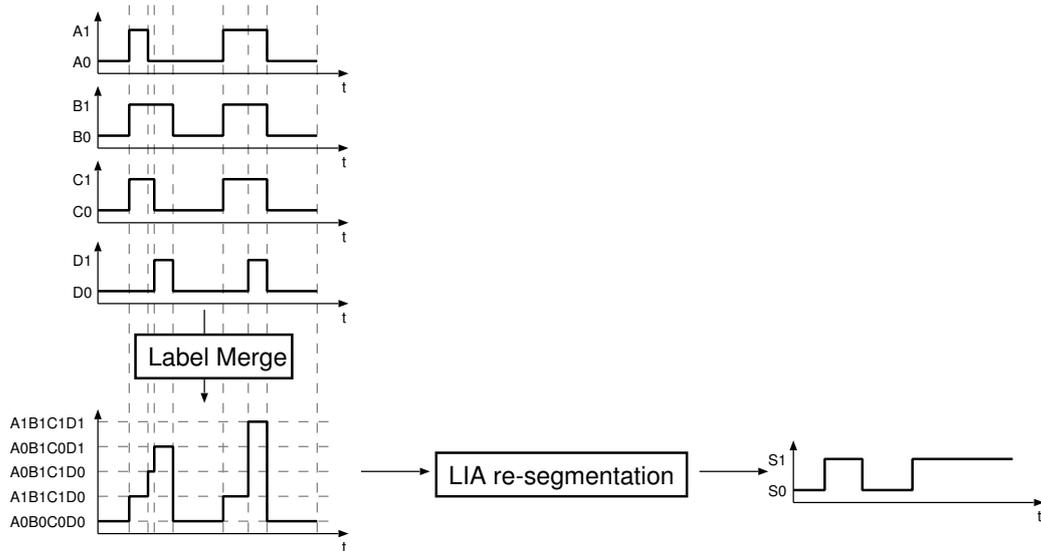
Fig. 7. *Example of a merging system*

4 Experiments and results

The experiments were carried out in the framework of the NIST-RT'03S speaker diarization evaluation on American broadcast news [4].

4.1 Development and evaluation corpora

Following the NIST-RT'03S evaluation campaign, two corpora are available for the speaker diarization task. One of them is used for the development of the systems, which are validated on the second one during a blind evaluation. The development corpus is extracted from the HUB-4 evaluation campaign corpus. It is composed of six broadcast news shows of about 10 minutes each, recorded in 1998 from channels MNB, CNN, NBC, PRI, VOA and ABC. The evaluation corpus is composed of three 30-minute shows recorded in 2001 from channels PRI, VOA and MNB, each containing between 10 and 27 speakers.

In this paper, these development and evaluation corpora are named respectively *RT'03S-Dev* and *RT'03S-Eva*. Two additional corpora are used during the experiments. Both of them are derived from *RT'03S-Dev* and *RT'03S-Eva* by discarding all the advertisement portions manually before being processed [2]. They are named *ELISA-Dev* (derived from *RT'03S-Dev*) and *ELISA-Eva* (derived from *RT'03S-Eva*) and serve the same role as the

---

[2] Advertisements, present in the audio documents, are never scored. Nevertheless, their presence during the segmentation process may disturb the systems since they involve additional speakers, entirely unrelevant in the output segmentation.

original corpora, *i.e.* system development and evaluation purposes. The use of these additional corpora during experiments may explain that some results presented in this paper do not correspond exactly to the official NIST-RT'03S results.

In order to evaluate the accuracy of the acoustic macro-class segmentation, a reference segmentation including the different targeted acoustic class (speech/non-speech, gender labels, and telephone/non telephone speech) was necessary. Since NIST does not provide any official reference for the bandwidth classification, the authors have marked their own. Both the boundaries and labels were manually identified. This reference segmentation will be referred to as *Hand S/NS-Gender-T/NT* later in this paper.

Moreover, it is worth noting that due to the small size of the different corpora, all the results presented in this paper have to be considered with caution.

4.2 Evaluation metric

The speaker diarization performance is evaluated by comparing the hypothesis segmentation, given by the system, with the reference segmentation provided by NIST. This reference segmentation was generated by hand according to a set of rules described in [4,34].

The evaluation metric is based on the NIST speaker diarization metric defined in the NIST-RT'03S evaluation plan [4]. It is called the diarization metric, and expressed in terms of Diarization Error Rate (DER). It takes three kinds of error into account (named SE, $MisE$, FaE respectively in the next sections):

- A speaker error defined below (SE).
- A missed speaker error relative to a misclassification of speech segments as non-speech segments ($MisE$).
- A false alarm speaker error relative to a misclassification of non-speech segments as speech segments (FaE).

To compute the speaker error, the scoring algorithm optimally maps the reference speakers to the hypothesis speakers. Each reference speaker is mapped onto one hypothesis speaker at most and conversely each hypothesis speaker is mapped onto one reference speaker at most. The mapping maximizes the overlap in duration between all pairs of reference and hypothesis speakers. The speaker error is finally expressed as the duration of non-matching zones between reference and hypothesis segments.

Concerning the gender- and bandwidth-misclassification errors, they are measured at a frame level by comparing the hypothesis classification with the reference segmentation proposed by the authors *Hand S/NS-Gender-T/NT*.

4.3 Acoustic macro-class segmentation experiments

This section presents the evaluation protocol used to measure the impact of the acoustic macro-class segmentation when combined with speaker diarization and discusses the experimental results obtained in this framework. Different levels of acoustic segmentation granularity are evaluated on both speaker diarization systems:

- Speech/non-speech classification only (*S/NS*). This segmentation corresponds to the first level of the acoustic macro-class segmentation described in section 2.1;
- segmentation based on speech/non-speech and gender detection (*S/NS-Gender*). This segmentation is obtained by merging all the labels *GS-XX*, *GSM-XX*, *GDS-XX* and *GT-XX* yielded by the acoustic macro-class segmentation (see figure 2) in a single *XX* label where *XX* represents either *Ma* or *Fe*.
- segmentation based on speech/non-speech, gender and telephone/non-telephone speech detection (*S/NS-Gender-T/NT*). *NT* segmentation is obtained by merging all the *GS-XX*, *GSM-XX*, and *GDS-XX* (see figure 2) in a single *NT-XX* label where *XX* represents either *Ma* or *Fe*.
- segmentation based on speech/non-speech, gender and telephone/clean speech/speech over music/degraded speech (*S/NS-Gender-T/S/MS/DS*). In this segmentation, all the labels yielded by the third level of the acoustic macro-class segmentation system are used (see figure 2).

For comparison purposes, speaker diarization results based on the reference acoustic macro-class segmentation, *Hand S/NS-Gender-T/NT*, are also presented.

4.3.1 Intrinsic performance of acoustic macro-class segmentation

Table 3 provides the performance of the a-prior acoustic macro-class segmentation on both *RT'03S-Dev* and *RT'03S-Eva* corpora. Some details about the amount of data for each targeted class are reported in table 4.
The speech/non-speech segmentation error is around 4.9% (in terms of duration) compared to 4.4% for the best system during the NIST-RT'03S evaluation campaign [35]. The gender detection error goes from 1.5% for the *RT'03S-Dev* set at 5.5% for the *RT'03S-Eva* set. As said in the description

of the corpora, the reference segmentation provided by NIST does not include telephone/non telephone information. Therefore, the accuracy of the acoustic segmentation system for the telephone and non-telephone classification is evaluated using reference boundaries marked by the authors (*Hand S/NS-Gender-T/NT*): less than 0.1% for the *RT'03S-Dev* corpus and 3% for *RT'03S-Eva*.

4.3.2 Performance of speaker diarization

This section presents the experimental results obtained when applying different levels of acoustic macro-class segmentation prior to the speaker diarization systems (*integrated* and *step-by-step* methods). Experiments are conducted on *ELISA-Dev* and *ELISA-Eva* corpora.

Table 5 provides the results obtained individually by each speaker diarization system before applying the re-segmentation step described in section 2.4 whereas table 6 provides the results obtained after the re-segmentation step. Three kinds of observation may be pointed out through these results, expressed in terms of missed speaker error rate (MiE), false alarm speaker error rate (FaE), speaker error rate (SE) and diarization error rate (DER):

- (a) concerning the corpora (*ELISA-Dev* and *ELISA-Eva*), a large variation in terms of performance may be observed between the speaker diarization systems depending on the corpus used. Indeed, the performance of the *integrated* system drastically decreases on *ELISA-Eva* corpus compared with *ELISA-Dev* (*e.g.* from 14.8% to 27.3% for *S/NS-Gender-T/NT* acoustic segmentation) while the *step-by-step* system performance remains quite steady whatever the corpus used.
- (b) concerning the acoustic macro-class segmentation, a gain in performance[3] for both speaker diarization systems can be observed when they are combined with manual acoustic segmentation. Similarly, a large improvement of the *integrated* approach results is obtained with the speech/non-speech, gender and telephone/non-telephone segmentations (*S/NS-Gender-T/NT*), especially on *ELISA-Eva* corpus, without (from 26.9% to 18.1%) and with (from 26.5% to 14.1%) the re-segmentation phase. On *ELISA-Dev* corpus, this improvement is more visible after the re-segmentation phase (from 15.5% to 12.8%) than before the re-segmentation for which only a small drop is observed (from 15.4% to 15.1%). On the other hand, even if some improvement can be noticed for the *step-by-step* system, the gain is minor. In fact, it is really visible after the re-segmentation phase (from 18.8% to 17.4% on *ELISA-Dev* and from 15.4% to 13.7% on *ELISA-Eva*). Finally, no improvement is seen (and in some cases, even a performance

---

[3] Best DER are reached using manual macro-class acoustic segmentation. Nevertheless, this is due to the MiE and FaE rates reduced to 0.

loss occurs) when the most detailed acoustic segmentation (S/NS-$Gender$-$T/S/MS/DS$) is involved. It can be noticed that this loss of performance becomes especially big without the re-segmentation phase.

- (c) applying the re-segmentation step leads to the best performance in most of the cases. This demonstrates its interest while coupled with both speaker diarization strategies.

Comparing all the different levels of segmentation granularity (note (b)), the S/NS-$Gender$-T/NT segmentation seems the most helpful for the speaker diarization task, especially for the *integrated* approach. This point is particularly visible for the *ELISA-Eva* corpus for which 20% of speech time (shared among 2 shows over the 3 available in the corpus) is telephone speech against 7.7% only for *ELISA-Dev* corpus (mainly present in 1 show over the 6 available).

The difference of behaviors in terms of performance (note (b)) between the two speaker diarization systems may be directly linked to the strategies involved for each of them. It seems reasonable that the *step-by-step* approach especially the speaker turn detection step intrinsically behaves as an acoustic class segmentation system, detecting speaker turns as well as acoustic event changes before the clustering phase. In this sense, the a-priori acoustic macro-class segmentation becomes useless to improve performance. Obviously this is not true for speech/non-speech detection since the speaker turn detection phase cannot discard non-speech segments automatically without additional processing.

Unexpectedly, the most detailed segmentation, S/NS-$Gender$-$T/S/MS/DS$, does not lead to performance gain and may conversely degrade it in most of the cases. This can be explained by the fact that some speakers may be present under different acoustic classes (speech over music followed by speech only, classical for news presenters, or in both clean and degraded speech classes depending on the location of interviews for instance). Since speaker diarization systems are applied independently on each acoustic class, the same speaker may be split under different labels, leading to an increase in speaker error rates. In the same way, increasing the number of acoustic classes induces much smaller segments, which may disturb speaker diarization systems. However, these effects are partially overcome thanks to the re-segmentation phase, which may explain that the loss of performance due to S/NS-$Gender$-$T/S/MS/DS$ segmentation is minor after applying the re-segmentation phase.

Finally, combining both speaker diarization systems with manual acoustic segmentation outperforms all the automatic ones. However, since the diarization error rate takes both speaker and speech/non-speech error rates into account, the results cannot be compared directly in this case (manual segmentation does not yield any speech/non-speech error rate). Regarding speaker error rate only, the best speaker diarization system (after re-segmentation) based on an auto-

matic acoustic segmentation on *ELISA-Dev* corpus gets 7.6% against 9.2% for manual segmentation. As a result, the speaker diarization system based on an automatic segmentation outperforms (from a pure speaker diarization point of view) the one based on a manual segmentation. In fact, the analysis of the results showed that some segmentation errors, due to some segments falsely split into two different classes (telephone/non telephone for instance) by the automatic acoustic segmentation system, may be automatically corrected by the re-segmentation step.

4.4   NIST-RT'03S results

This section presents the results obtained during the NIST-RT'03S evaluation campaign, on the *RT'03S-Dev* and *RT'03S-Eva* corpora, for both speaker diarization approaches as well as the "fusion" systems, described in sections 3.1 and 3.2. It can be underlined that:

- the "merging" strategy-based system (ELISA1), submitted as ELISA primary system, obtained the second lower diarization error rate compared to the other NIST-RT'03S-participant primary systems [35].
- the "hybridization" strategy-based system (ELISA2) (*i.e.* the CLIPS system followed by the re-segmentation phase), submitted as a secondary ELISA system, outperformed the best primary system [36,37] and obtained the lowest speaker diarization error rate.

Table 7 summarizes the performance achieved by the different proposed systems during the NIST-RT'03S. It shows that:

- even though the five systems are based on the same acoustic macro-class segmentation, the Miss Speech and False Alarm Speech errors are different. This is due to the LIA and ELISA system behavior, which work at 0.2 second block level [4] (all the segment boundaries are aligned on a 0.2 second scale) whereas CLIPS system works at a frame level. It gives small differences in the boundary positions of the segments.
- The LIA E-HMM based primary system (LIA1) improves performance compared with the CLIPS classical approach (CLIPS1) (16.9% DER compared to 19.3%). But the second LIA system (LIA2) based upon a linear model adaptation descibed in [9] obtained only 24.7% DER. This difference in terms of performance illustrates the difficulty of adapting a large statistical model in borderline conditions (only few seconds of adaptation data).
- the *integrated* (E-HMM) method may clearly benefit from some better segment boundaries and longer segments issued from a classical turn detection

---

[4]   see sections 2.3.2 and 2.4 for details on the 0.3 second block level (for segmentation phase) and 0.2 second (for re-segmentation) signification.

approach like the CLIPS one, as demonstrated by the a large gain of performance (from 16.9% to 12.9% DER) reached by the ELISA2 system. Indeed the re-segmentation phase improves the accuracy of the CLIPS diarization and it allows to reduce the diarization error by 33% (relative).

- the strategy involved in ELISA1 system performs better than ELISA2 over two recordings while a drastic loss is observed on the last recording. The loss on that particular recording is a good example of the limitation of the merging technique explained in section 3.2: one of the systems disagreed with the others. This resulted in too many speakers detected and, most important, in the split of a long true speaker into two hypothetic speakers, involving a large error rate. This observation illustrates the remark done in section 4.1, in which it is noted that some caution has to be taken with the result discussion due to the small size of the different corpora. In other words, the generalization problem is underlined here with only three different shows available for testing.

- for the CLIPS system, complementary experiments showed that automatically estimating the number of speakers during the clustering process generates only about 4% more of absolute diarization error than the optimal number of speakers. The CLIPS algorithm missed only 7% of the real speakers involved in the files (4 speakers out of 57 total speakers). It is important to note here that we call "optimal number of speakers" the number of speakers that minimizes the diarization error and not the real number of speakers involved in the conversation. The optimal number is usually smaller than the real number due to the fact that there are a lot of speakers, especially in broadcast news data, that do not speak enough to train a reliable statistical model (e.g: 4 seconds during a 30-minute file). To illustrate this point, table 8 presents the speaker diarization error using respectively the optimal, the estimated and the real number of speakers obtained on two speech corpora.

Moreover, we observed that the label merging method of the four systems generates about 150 potential speakers per show. This speakers correspond generally to:

- potential speakers that have a large amount of data assigned (>10 seconds). These speakers can be considered as correct hypothesized speakers;

- potential speakers generated by few systems, for example the speakers associated with only one short segment ($\leq$10 seconds). These hypothesized speakers can be suppressed (the weight of these speakers on the final scoring is marginal);

- potential speakers that have a small amount of data scattered between multiple small segments and that can be considered as zones of indecision.

We observed also that after the first iteration of the re-segmentation, the number of speakers is already drastically reduced (from 150 to about 50) since speakers associated with indecision zones do not catch any data during

the Viterbi decoding and are automatically removed. However, the merging strategy cannot generally solve the wrong behavior of initial systems that can split a "true" speaker into two hypothesized speakers, each tied to a long segment. Suppose all systems agree on a long segment except for one which splits this segment into two parts. This would produce two potential speakers (associated with long duration segments) after the merging phase and since we do not do any clustering before re-segmentation, we have generally a "true" speaker split into two hypothesized speakers.

5 Ongoing works

5.1 Application to other data

Though this paper was mainly dedicated to speaker diarization experiments on broadcast news data, our speaker diarization systems were successfully applied to other data types during the last NIST evaluations. Table 9 presents a summary of our results obtained since 2002 on different kinds of data. In 2002, beside the diarization of broadcast news documents, two other tasks were proposed on telephone speech conversations and on meeting room recordings. Performance shown on the first line of the table illustrates the increasing difficulty of the tasks. For telephone conversations, only two persons are involved and there are few overlapping segments. For broadcast news, there are obviously more speakers on the audio documents, but this is mostly prepared speech with a large part of "studio quality" voice. The hardest task definitely corresponds to meeting data with very spontaneous speech, recovering voices, disfluencies, distant speakers (in case of table microphones) and background noise. The second line shows the best performance obtained in spring 2003 on broadcast news data with the system described in this paper, and illustrates the progress made from 2002 to spring 2003 on this data. During the NIST 2004 spring rich transcription evaluation [21], the novelty was that we had to process multiple speech channels coming from multiple sensors located in a smart meeting space. We proposed a very straightforward strategy to merge the multiple channel segmentation outputs and obtained the best speaker diarization performance for this task [21].

5.2 Integration of a-priori knowledge

One of the main assumptions in most of the papers [32,9] concerning the speaker diarization task is that no a-priori information is available on the test data. This means for instance, that the number and the identity of the speakers

involved in conversation are not known. A consequence of this hypothesis is that no reference speaker data is supposed to be available before segmenting an audio signal for instance.

However, this limitation may not be so necessary [38] for some applications and conditions for which we can reasonably hope to have a-priori information. For instance, the type of conversation is generally known (broadcast news, telephone or meeting conversation), which gives us information on speech quality and average speaker turn length. We may also know about the real number of speakers involved in conversation. For instance, we know there are two speakers in telephone conversation while a list of participants might be available for meeting data. In some cases reference data might be available for the speakers involved in the conversation. For example, we could ask every participant to introduce themselves at the beginning of meetings. A synthesis of what kind of information we might expect for each type of audio document is presented in Table 10.

Some results concerning the knowledge of the real number of speakers were already presented in section 2.2.3. From those results only it was difficult to conclude if the knowledge of the number of speakers is useful information or not since the conclusion is different for the two speech corpora.

In the case of broadcast news data we can easily obtain reference data of one particular speaker. This speaker is the news show presenter. From previously broadcasted shows, it is thus possible to obtain enough data to train a presenter speaker model directly by EM. We have shown in [38] that using a simple speaker tracking system for this particular speaker, up to 3% absolute diarization error reduction can be obtained (experiments done on the ESTER [5] radio broadcast corpus, see [38] for more experimental details).

The possibility of having reference data for all speakers is specific to telephone and meeting data. The main interest in having data available for all speakers, other than a significant error reduction (up to 10% absolute for the *RT03-Eva* corpus, see [38] for experimental details) is the execution time consideration. Our *step-by-step* speaker diarization approach takes about four times real-time for a 30 minutes file. When reference data is available for all speakers, the speaker diarization is a simple assignment of every segment to the most likely speaker. In this case the speaker diarization takes about 0.1 times real-time. Every speaker model is derived from a background model using the few seconds available for each speaker.

In conclusion, with the current modeling technique including in our systems, the real number of speakers is useful only as a reference number for the esti-

---

[5] www.afcp-parole.org/ester/

mated number of speakers, whereas reference data for the speakers involved in conversation is always useful. Further experiments should be done on telephone and meeting data concerning the case when reference data is available for all speakers.

6   Conclusion

This paper summarizes the collaboration of the LIA and CLIPS laboratories, two members of the ELISA consortium, in the area of speaker diarization. The work presented in this paper was done in the framework of the NIST Rich Transcription 2003 spring evaluation campaign (NIST-RT'03S) and addressed two main points.

Firstly, two main approaches for speaker diarization were proposed and compared. The first one relies on a classical strategy based on speaker-turn detection followed by a clustering process, while the second one relies on an *integrated* strategy where segment boundaries and speaker tying of the segments are extracted simultaneously and challenged during the whole process. The *integrated* method (E-HMM) shows a higher modeling power (16.9% diarization error compared to 19.3% for the *step-by-step* approach). Nevertheless, the classical *step-by-step* approach seems to obtain more consistent results across the files and conditions than the *integrated* one. Despite the differences between the two approaches, the results obtained during the NIST 2003 spring evaluation showed the interest of using both techniques. This was confirmed by the results obtained using a fusion of both systems where the *integrated* approach is applied after the CLIPS *step-by-step* segmentation system. The fusion system obtained an error rate of 12.9% as compared to 19.3% for the CLIPS system used on its own. The fusion system also showed a relative 33% error-reduction compared to the performance of the *integrated* system taken alone (from 16.9% to 12.9%). The *integrated* (E-HMM) method clearly benefits from the better segment boundaries and longer segments issued from the classical CLIPS approach. This fusion system achieved the lowest speaker diarization error rate during the NIST-RT'03S evaluation campaign. More investigation is needed for a better understanding of the nature of the errors made by the systems, which is not a trivial task as the speaker diarization performance metrics is complicated.

The second main issue addressed in this work concerns the nature of the audio documents to segment. This paper focuses on the case of audio broadcast news documents. An acoustic macro-class segmentation was proposed, as a-prior step for speaker segmentation systems. The speaker segmentation system is run independently on each acoustic sub-class and the resulting segmentations are merged thanks to a re-segmentation algorithm (in this paper, the

27

re-segmentation process consists of one iteration of the *integrated* E-HMM algorithm). For a speech/non-speech, gender and bandwidth (studio/telephone speech) pre-segmentation, a significant gain was observed in the case of the *integrated* approach. A slight gain was also observed for the *step-by-step* approach, which seems more robust to channel or environment variations. Moreover, finer macro-class segmentation (including speech over music detection) led to a loss in performance, partially due to the assumption (fixed during the speaker segmentation process) that the same speaker could not appear in more than one acoustic macro-class.

Finally, some ongoing work has been presented, which demonstrates that the approaches proposed are able to deal with new types of data like meeting room recordings. Indeed, multiple microphones are often available in the case of meeting rooms and taking these multiple (and low quality) recordings of the same conversation into account constitutes a new challenge for speaker diarization. The systems presented in this paper were adapted to the meeting task and obtained during the NIST-RT'03S campaign a state-of-the-art result with 22.8% diarization error rate. A strategy for fusing segmentations issued from several microphones was also proposed but no significative win was observed compared with results obtained using the best microphone. The interest of a-priori knowledge concerning the potential number of speakers nor the speakers was also presented. Particularly, using knowledge from well-known speakers authorizes a 3% absolute win during experiments on ESTER database and knowing all the potential speakers allows to speed up very significantly the diarization process. This preliminary work on using available apriori information opens up interesting possibilities for further work.

References

[1] DARPA speech recognition evaluation workshop, `http://www.nist.gov/speech/publications/`.

[2] D. Y. Kim, G. Evermann, T. Hain, D. Mrva, S. Tranter, P. Wang, L. Woodland, Recent advances in broadcast news transcription, in: Automatic Speech Recognition and Understanding, IEEE, ASRU 2003, St. Thomas, U.S. Virgin Islands, 2003, pp. 105–110.

[3] L. Nguyen, B. Xiang, Light supervision in acoustic model training, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2004), Montreal, Canada, 2004.

[4] NIST, The rich transcription spring 2003 (RT-03S) evaluation plan, `http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf`, (Version 4, Updated 02/25/2003) (February 2003).

[5] M.-H. Siu, R. Rohlicek, H. Gish, An unsupervised, sequential learning algorithm for segmentation of speech waveforms with multi speakers, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 92), Vol. 2, San Fransisco, CA, 1992, pp. 189–192.

[6] L. Wilcox, F. Chen, D. Kimber, V. Balasubramanian, Segmentation of speech using speaker identification, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 94), Adelaide, Austrailia, 1994, pp. 161–164.

[7] M. Siegler, U. Jain, B. Raj, R. Stern, Automatic segmentation and clustering of broadcast news audio, in: The DARPA Speech Recognition Workshop, Westfields, Chantilly, Virginia, 1997.

[8] S. Chen, P. Gopalakrishnan, Speaker, environment and channel change detection and clustering via the bayesian information criterion, in: DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, 1998.

[9] S. Meignier, J.-F. Bonastre, S. Igounet, E-HMM approach for learning and adapting sound models for speaker indexing, in: 2001 : a Speaker Odyssey. The Speaker Recognition Workshop, Chania, Crete, 2001, pp. 175–180.

[10] J. Ajmera, C. Wooters, A robust speaker clustering algorithm, in: Automatic Speech Recognition and Understanding, IEEE, ASRU 2003, St. Thomas, U.S. Virgin Islands, 2003, pp. 411–416.

[11] D. A. Reynolds, R. B. Dunm, J. J. Laughlin, The Lincoln speaker recognition system: NIST EVAL2000, in: Proceedings of International Conference on Spoken Language Processing (ICSLP 2000), Vol. 2, Beijing, China, 2000, pp. 470–473.

[12] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre, The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2004), Montreal, Canada, 2004.

[13] NIST, The NIST 2001 speaker recognition evaluation plan, `http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf` (March 2001).

[14] NIST, The NIST Year 2002 speaker recognition evaluation plan, `http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrec-evalplan-v60.pdf` (February 2002).

[15] NIST, Spring 2004 (RT-04S) rich transcription meeting recognition evaluation plan, `http://www.itl.nist.gov/iad/894.01/tests/rt/rt2004/spring/documents/rt04s-meeting-eval-plan-v1.pdf` (February 2004).

[16] A. Smeaton, W. Kraaij, P. Over, TRECVID 2003 - an introduction, in: 12th Text Retrieval Conference, 2003.

[17] G. Quénot, D. Moraru, L. Besacier, P. Mulhem, Clips-imag at trec-11 : Experiments in video retrieval, in: TREC 2002, Gaithersburg, MD, USA, 2002.

[18] G. Quénot, D. Moraru, L. Besacier, Clips at TRECvid: Shot boundary detection and feature detection, in: TREC 2003, Gaithersburg, MD, USA, 2003.

[19] ELISA, The ELISA systems for the NIST 99 evaluation in speaker detection and tracking, Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3) (2000) 143–153.

[20] I. Magrin-Chagnolleau, G. Gravier, R. Blouet, for the ELISA consortium, Overview of the ELISA consortium research activities, in: 2001 : a Speaker Odyssey. The Speaker Recognition Workshop, Chania, Crete, 2001, pp. 67–72.

[21] C. Fredouille, D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, The NIST 2004 spring rich transcription evaluation : two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation, in: RT2004 Spring Meeting Recognition Workshop, 2004, p. 5.

[22] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, Y. Magrin-Chagnolleau, The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Vol. II, Hong Kong, 2003, pp. 89–92.

[23] P. Delacourt, C. J. Welkens, DISTBIC: A speaker based segmentation for audio data indexing, Speech Communication 32 (2000) 111–126.

[24] T. Hain, P. Woodland, Segmentation and classification of broadcast news audio, in: Proceedings of International Conference on Spoken Language Processing (ICSLP 98), Sydney, Australia, 1998.

[25] P. Woodland, The development of the HTK broadcast news transcription system: An overview, Speech Communication 37 (1-2) (2002) 291–299.

[26] J.-L. Gauvain, L. Lamel, G. Adda, The LIMSI broadcast news transcription system, Speech Communication 37 (1-2) (2002) 89–108.

[27] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (2) (1978) 461–464.

[28] S. Meignier, J.-F. Bonastre, C. Fredouille, T. Merlin, Evolutive HMM for speaker tracking system, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2000), Istanbul, Turkey, 2000, pp. 1177–1180.

[29] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted gaussian mixture models, Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3) (2000) 19–41.

[30] L. Wilcox, D. Kimber, F. Chen, Audio indexing using speaker identification, in: Proceedings SPIE Conference on Automatic Systems for the Inspection and Identification of Humans, San Diego, CA, 1994, pp. 149–157.

[31] J.-L. Gauvain, L. Lamel, G. Adda, Audio partitioning and transcription for broadcast data indexation, Multimedia Tools and Applications (2001) 187–200.

[32] A. Adami, S. S. Kajarekar, H. Hermansky, A new speaker change detection method for two-speaker segmentation, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2002), Vol. IV, 2002, pp. 3908–3911.

[33] J.-L. Gauvain, C. H. Lee, Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains, in: IEEE Transactions on Speech and Audio Processing, Vol. 22, 1994, pp. 291–298.

[34] NIST, Reference data cookbook for Who Spoke when diarization task, `http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/ref-cookbook-v2_4.pdf`, v2.4 (2003).

[35] NIST, Who Spoke When Speaker-ID and Speaker-Type Metadata diarization, `http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/RT03s_Diarization.pdf` (May 2003).

[36] P. Nguyen, J.-C. Junqua, PSTL's speaker diarization, `http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/panasonic-spkr-rt03.pdf` (May 2003).

[37] Y. Moh, P. Nguyen, J.-C. Junqua, Towards domain independent speaker clustering, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, 2003.

[38] D. Moraru, L. Besacier, E. Castelli, Using a priori information for speaker diarization, in: 2004 : A Speaker Odyssey. The Speaker Recognition Workshop, Toledo, Spain, 2004, pp. 355–362.

Table 1
Increasing difficulty of the tasks. Best results for the speaker diarization task in the
NIST'02 speaker recognition evaluation.

Task	Telephone	Broadcast News	Meeting
Diarization error rate	5.7	26.4	30.1
Problems involved	p1 (but with a fixed number of speakers)	p1+p2	p1+p2+p3

Table 2
Description of the different corpora, in terms of number of a shows, size average,
and speaker number average. (* As advertisements are not manually transcribed,
the exact number of speakers is unknown)

Corpus	Number of shows	Size average (in sec.)	Speaker Nb average
RT'03S-Dev	6	568	>13*
RT'03S-Eva	3	1534	>20
Elisa-Dev	6	574	13
Elisa-Eva	3	1773	20

Table 3
Acoustic macro-class segmentation error rates on the *RT'03S-Dev* and *RT'03S-Eva*
sets.

Corpus	Misclassification error rate (in %)			
	Speech	Non-speech	Gender	Tel./Non Tel.
RT'03S-Dev	2.3	2.2	1.5	0.09
RT'03S-Eva	1.1	3.8	5.5	3.0

Table 4
Amount of data for each targeted acoustic class: speech/non speech classes, female
and male speech classes, telephone and non-telephone speech class.

Corpus	Data amount (in sec.) of each acoustic class in the corpus					
	speech	non-speech	female	male	Telephone	non-telephone
RT'03S-Dev	3090	321	730	2360	220	2870
RT'03S-Eva	4127	478	1271	2856	530	3597

Table 5

Error rates, expressed in terms of missed speaker (MiE), false alarm speaker (FaE), speaker (SE) and diarization speaker (DER) error rates (in %), obtained by each speaker diarization system before applying the re-segmentation step when combined with different levels of acoustic macro-class segmentation. Experiments conducted on *ELISA-Dev* and *ELISA-Eva* corpora.

| | *Step-by-step* system | | | | | | | |
| | *ELISA-Dev* | | | | *ELISA-Eva* | | | |
Acoustic segmentation	MiE	FaE	SE	DER	MiE	FaE	SE	DER
Hand S/NS-Gender-T/NT	*0.0*	*0.0*	*14.0*	*14.0*	*0.0*	*0.0*	*10.2*	*10.2*
S/NS	2.8	2.4	14.5	19.7	2.1	3.0	12.2	**17.3**
S/NS-Gender	2.8	2.4	13.5	**18.7**	2.1	3.0	13.6	18.7
S/NS-Gender-T/NT	2.8	2.4	13.9	19.1	2.1	3.0	13.3	18.4
S/NS-Gender-T/S/MS/DS	2.8	2.4	19.5	24.7	2.1	3.0	22.5	27.6

| | *Integrated* system | | | | | | | |
| | *ELISA-Dev* | | | | *ELISA-Eva* | | | |
Acoustic segmentation	MiE	FaE	SE	DER	MiE	FaE	SE	DER
Hand S/NS-Gender-T/NT	*0.0*	*0.0*	*10.7*	*10.7*	*0.0*	*0.0*	*12.0*	*12.0*
S/NS	2.8	2.4	10.2	15.4	2.1	3.0	21.8	26.9
S/NS-Gender	2.8	2.4	9.6	**14.8**	2.1	3.0	22.2	27.3
S/NS-Gender-T/NT	2.8	2.4	9.9	15.1	2.1	3.0	13.0	**18.1**
S/NS-Gender-T/S/MS/DS	2.8	2.4	18.0	23.2	2.1	3.0	23.0	28.1

Table 6
Error rates, expressed in terms of missed speaker (MiE), false alarm speaker (FaE), speaker (SE) and diarization speaker (DER) error rates (in %), obtained by each speaker diarization system after applying the re-segmentation step when combined with different levels of acoustic macro-class segmentation. Experiments conducted on *ELISA-Dev* and *ELISA-Eva* corpora.

| | **Step-by-step** system | | | | | | | |
| | *ELISA-Dev* | | | | *ELISA-Eva* | | | |
Acoustic segmentation	MiE	FaE	SE	DER	MiE	FaE	SE	DER
Hand S/NS-Gender-T/NT	*0.0*	*0.0*	*13.7*	*13.7*	*0.0*	*0.0*	*10.5*	*10.5*
S/NS	2.8	2.4	13.6	18.8	2.1	3.0	10.3	15.4
S/NS-Gender	2.8	2.4	12.5	17.7	2.1	3.0	10.0	15.1
S/NS-Gender-T/NT	2.8	2.4	12.2	**17.4**	2.1	3.0	8.6	**13.7**
S/NS-Gender-T/S/MS/DS	2.8	2.4	12.3	17.5	2.1	3.0	9.4	14.5

| | **Integrated** system | | | | | | | |
| | *ELISA-Dev* | | | | *ELISA-Eva* | | | |
Acoustic segmentation	MiE	FaE	SE	DER	MiE	FaE	SE	DER
Hand S/NS-Gender-T/NT	*0.0*	*0.0*	*9.2*	*9.2*	*0.0*	*0.0*	*10.8*	*10.8*
S/NS	2.8	2.4	10.3	15.5	2.1	3.0	21.4	26.5
S/NS-Gender	2.8	2.4	7.8	13.0	2.1	3.0	19.8	24.9
S/NS-Gender-T/NT	2.8	2.4	7.6	**12.8**	2.1	3.0	9.0	**14.1**
S/NS-Gender-T/S/MS/DS	2.8	2.4	9.8	15.0	2.1	3.0	9.0	**14.1**

Table 7
Experimental results on NIST-RT'03S data.

Different error rates (in %)	Miss Speech	FA Speech	Speaker	Diarization
CLIPS (CLIPS1)	2.0	2.9	14.3	19.3
CLIPS + re-seg. (ELISA2)	1.1	3.8	8.0	12.9
LIA (LIA1)	1.1	3.8	12.0	16.9
LIA variant (LIA2)	1.1	3.8	19.8	24.7
merging strategy (ELISA1)	1.1	3.8	9.3	14.2

Table 8

Speaker diarization error rate using different estimates of the number of speakers. The error rate is extracted from the CLIPS NIST-RT'03S results. N_{Opt}: optimal number, the number of speakers that gives the minimal error (total=46 on RT03Dev). N_{Est}: estimated number of speakers (total=47 on *RT'03S-Dev*). N_{Real} real number of speakers (total=69 on *RT'03S-Dev*).

Corpus	N_{Opt}	N_{Est}	N_{Real}
RT'03S-Dev	14.5	19.7	24.8
RT'03S-Eva	14.0	17.1	16.3

Table 9

ELISA Results since 2002 (diarization error rate) given various corpora

Corpus / Year	Telephone	Broadcast News	Meeting (head mic.)	Meeting (table mic.)
2002	5.7	30.3	34.7	36.9
spring 2003	X	12.9	X	X
spring 2004	X	X	X	22.4

Table 10

a-priori information available for each audio document type

Information	Telephone	Meeting	Broadcast News
N Speakers	2	Possibly Known	Unknown
Reference Data Available	Possibly (2 Spk)	Possibly (1:N Spk)	Possibly (Few Spk Only)

35