SPEAKER DIARIZATION IN THE ELISA CONSORTIUM OVER THE LAST 4 YEARS

Daniel Moraru, Laurent Besacier, Sylvain Meignier, Corinne Fredouille, J.-F Bonastre

SPEAKER DIARIZATION IN THE ELISA CONSORTIUM OVER THE LAST 4 YEARS

D. Moraru, L. Besacier

CLIPS - IMAG
(UJF & CNRS)
Grenoble (France)
(laurent.besacier,daniel.moraru)@imag.fr

S. Meignier[*]*, C. Fredouille, J.-F. Bonastre*

LIA/CERI Université d'Avignon,
Avignon (France)
(corinne.fredouille,jfb)@lia.univ-avignon.fr
sylvain.meignier@univ-lemans.fr

ABSTRACT

This paper summarizes the collaboration of the LIA and CLIPS laboratories, members of the ELISA consortium, along the last 4 year NIST speaker diarization system evaluation campaigns. In this context, two individual approaches, quite different, have been developed individually by each lab, to respond to the specific task of speaker segmentation. The first one relies on a classical two-step speaker segmentation strategy, based on the detection of speaker turns followed by a clustering process, while the second one corresponds to an integrated strategy where both segment boundaries and speaker tying of the segments are extracted simultaneously and challenged during the whole process. From these two main methods, various strategies were investigated for the fusion of segmentation results.

Through the performance achieved along the different evaluation campaigns as well as the experience gained by the LIA and CLIPS labs in the speaker diarization task, a discussion about the overall work done in this evaluation context is drawn in this paper, proposing further investigation and progression.

1. INTRODUCTION

Since 1996, NIST has organized yearly speaker recognition evaluation campaigns, focusing on the automatic speaker detection task. In 2000, the evaluation of speaker segmentation systems was introduced as a new task. Also called speaker diarization, this task consists in segmenting a conversation involving multiple speakers into homogeneous parts which contain the voice of only one speaker, and in grouping together all the homogeneous segments that correspond to the same speaker.

In parallel, the progress made in broadcast news transcription, moves the focus on a new task, denoted "rich transcription", for which the semantic information is not the only element of interest. Indeed, acoustic based information (sounds, speech qualities, speaker information, ...), discourse based information (disfluencies, emotion, ...), as well as linguistic information (topic, named entities, ...) may also be used to enrich the transcription and to help for indexing audio documents. Speaker characteristics are obviously an important information in this context. For this reason, the speaker diarization system evaluation has joined in 2003 the Rich Transcription system evaluation campaign.

The LIA and CLIPS labs, members of the ELISA consortium, have participated since 2000 (only LIA in 2000 and 2001) in these evaluation campaigns [1, 2, 3]. Two main strategies have been proposed and improved over the last four evaluation campaigns.

One of the main characteristics of the NIST evaluation campaigns has been to propose different kinds of environment to evaluate the speaker diarization systems: telephone conversational speech, broadcast news shows as well as meeting room recordings. This paper presents the progression of the LIA and CLIPS systems in terms of speaker segmentation strategies as well as in terms of performance, according to the targeted environment. It is organized as follows: section 2 is dedicated to the LIA and CLIPS baseline speaker segmentation system description. Section 3 presents the evolution of these systems as well as further investigation to improve performance, based for instance on system fusion. Performance of the best ELISA systems over the last four years is provided in section 4. Regarding this performance, a discussion is proposed, underlining the issues raised over the last evaluation campaigns. Finally, section 5 concludes this paper and gives some perspectives.

2. BASELINE OF SPEAKER DIARIZATION SYSTEMS

Two different speaker segmentation systems are presented in this section. They have been developed individually by

[*]is at LIUM lab of the University of Le Mans since September 2004

the CLIPS and LIA labs in the framework of the ELISA consortium [1, 2, 3], using AMIRAL, the LIA Speaker Recognition system [4]. The CLIPS system relies on a classical two-step strategy. It involves a BIC (Bayesian Information Criterion) detector based strategy for speaker turn detection followed by a hierarchical clustering. This approach will be denoted as "step-by-step strategy" in the rest of this paper, as the information retrieved during the speaker turn detection is not questioned during the clustering phase. The second system developed by the LIA differs from the previous one by proposing an "Integrated" strategy, for which all the information is iteratively questioned along the segmentation process. It is based on a HMM modeling of the conversation and an iterative process which adds the speakers one by one.

2.1. CLIPS Step-by-Step approach

Approach overview
The CLIPS system is based on a BIC (Bayesian Information Criterion)[5, 6, 7] speaker change detector (Step one) followed by a hierarchical clustering (Step two). A BIC curve is extracted by computing a distance between two 1.75s adjacent windows that go along the signal. Mono-Gaussian models with diagonal covariance matrices are used to model the two windows. A threshold is then applied on the BIC curve to find the most likely speaker change points which correspond to the local maximums of the curve. Clustering starts first by training a 32 component GMM background model (with diagonal covariance matrices) on the entire test file maximizing a ML criterion thanks to a classical EM algorithm. Segment models are then trained using MAP adaptation of the background model (means only). Next, BIC distances are computed between segment models and the closest segments are merged at each step of the algorithm until N segments are left (corresponding to the N speakers in the conversation).

System specification
The signal is characterized by 16 Mel Cepstral features (MFCC) computed every 10ms on 20ms windows, augmented by the energy. No frame removal or any coefficient normalization is applied.

2.2. LIA integrated approach

Approach overview
The LIA system shows a different strategy, based on a Hidden Markov Modeling (HMM) of the conversation and an iterative process which adds the speakers one by one [8]. Each state of the HMM characterizes a speaker and the transitions model the changes between speakers. During the segmentation, 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 composed of four steps:

- Step 1-Initialization. A first "speaker" model is trained on the whole test utterance (it is rather a generic acoustic model than a given speaker model). The conversation is modeled by a one-state HMM and the whole signal is set to the initial "speaker".
- Step 2-Adding a new speaker. A new speaker model is trained using 3 seconds of test speech that maximize the likelihood ratio computed using the first model and a world model (learned using development data). A corresponding state is added to the previous HMM.
- Step 3-Adapting speaker models. First, all the speaker models are adapted, using a MAP approach, according to the current segmentation. Then, a Viterbi decoding is done and produces a new segmentation. The adaptation and decoding steps are performed while the segmentation differs between two successive "adaptation/decoding" phases.
- Step 4-Assessing the stop criterion. The likelihood of the previous solution and the likelihood of the last solution are computed using the last HMM model (for example, the solution with two speakers detected and the 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. Two heuristic criteria are added to the likelihood-based criterion:
  - The first one removes the new speaker if the length of its segments is less than 4 seconds. The 3 second segment used for the initialization of the speaker is then marked as unavailable for the speaker initialization (step 2). The process continues with the segmentation of the previous iteration.
  - The second one discards the previous speaker from the segmentation if the length of their segments is lower than the new 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.

System specification
The signal is characterized by linear Cepstral features (LFCC)[1]

[1]The number of parameters may differ according to the task context.

computed every 10 ms using a 20ms window, augmented by the energy. No frame removal or any coefficient normalization is applied. GMM with 128 components (diagonal covariance matrix) are used for the speakers and world / background models.

3. CONTEXT DEPENDENT SYSTEMS

3.1. Overview of evaluation campaigns

The speaker segmentation task (also named speaker diarization or "who spoke when" task in the NIST terminology) was introduced in the NIST speaker recognition system evaluation campaign in 2000, in addition to the classical tasks based on the speaker detection. The main difference between speaker segmentation and the other tasks lies in the unavailability of prior information concerning the speakers involved in the speech signal: neither speaker identities nor speaker numbers. This particular context obviously increases the difficulties of the speaker segmentation task.
Chronologically, from 2000 to 2002, this specific task was proposed during the NIST speaker recognition system evaluation campaigns [9, 10, 11]. Since information retrieved from the speaker segmentation of an audio document can be easily seen as a way of enriching the transcription of this same document, the speaker segmentation system evaluation has joined the Rich Transcription system evaluation campaigns, also organized by NIST, in 2003 [12, 13].
Along these evaluation campaigns, speaker segmentation systems were evaluated on different kinds of data:

- conversational telephone speech corpora, which normally involve two speakers, and one acoustic class only for the signal (telephone speech) for NIST-SpRec-2000, NIST-SpRec-2001, and NIST-SpRec-2002 evaluation campaigns;
- broadcast news shows, which may contain a large set of speakers, over various acoustic classes (studio speech, telephone speech, degraded speech, speech over music, music, ...) for NIST-SpRec-2002 and NIST-RT-2003 evaluation campaigns;
- meeting data, collected through (distant) table or head microphones, involving few speakers (compared with broadcast news data), but more spontaneous speech (disfluencies, voice overlapping, long silences, ...), and eventually multi-channel signals, for NIST-SpRec-2002 (mono-channel signal only) and NIST-RT-2004 (Spring) (mono- and multi-channel signals) evaluation campaigns.

These differences between corpora raise two main observations. First of all, the second main difficulty[2] of the speaker segmentation task strongly depends on the type of targeted data (from fixed number of speakers to "unlimited", from one acoustic class to various ones, mono-channel vs multi-channel, ...). Secondly, speaker segmentation system has to be adapted according to the type of processed data, in order to increase performance.

3.2. From baseline to context dependent systems

The CLIPS and LIA labs have participated in the speaker segmentation system evaluation campaigns since 2001 (LIA only for this year), in association with the ELISA consortium. Considering the baseline speaker segmentation systems described in the previous section, various evolutions have been proposed to cope with the different kinds of data and to increase system performance. This section will present the most important ones for each individual system and those concerning both of them.
It has to be noted that LIA initiated its participation in speaker segmentation evaluation campaigns, in 2000, in collaboration with the EURECOM Institute (France). The system proposed consisted in two steps: the first step relying on a BIC based speaker turn detection, developed by the EURECOM Institute, followed by a step, based on the LIA speaker verification algorithms for segment aggregation. Some results concerning this system are provided in section 4.

3.2.1. Step-by-Step approach

Two main evolutions were implemented in the CLIPS step-by-step approach. The first one relies on the parameterization step for which more filter banks were used to process broadcast news data during the NIST-RT-2003 evaluation campaigns (56 filter banks) compared with the NIST-SpRec-2002 (24 filter banks only). Experimental results show in the former case, a performance improvement in terms of speaker segmentation error rates.
Secondly, during the NIST-SpRec-2002 evaluation, the CLIPS work was mainly focused on telephone speech data (considered as the primary task of the evaluation). Therefore, the speaker segmentation system was constrained to find only two speakers during the speaker segmentation process, whatever the data processed. Obviously, this configuration was well-suited for conversational telephone speech, but unrealistic for broadcast news and meeting corpora, involving drastic speaker segmentation performance in 2002.
For NIST-RT-2003 in which the speaker segmentation task

[2] the first one being directly linked to the intrinsic constraint of the task: no prior information about speakers involved in the audio documents.

was dedicated to broadcast news data, the CLIPS lab investigated an original method for estimating the number of speakers, appearing in the audio documents. In this approach, the number of speakers in the conversation (N_{Sp}) is estimated using a penalized BIC (Bayesian Information Criterion). The number of speakers is constrained between 1 (if we are working on an isolated acoustic pre-segmentation class, see section 3.2.3 for more details) or 2 (if we are working on the entire audio file) and 25. The upper limit is related to the recording duration. The number of speakers (N_{sp}) is selected to maximize:

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

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 equal to 0.6. 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 $L(X|M)$.
For NIST-RT-2004 evaluation campaign, no particular evolution was made on the system compared with the one used for broadcast news corpus in 2003.

3.2.2. Integrated approach

Few evolutions were implemented on the LIA integrated system, depending on the data processed. First of all, as this system relies on a UBM world model for speaker model adaptation, suitable data were chosen to estimate it (Switchboard phase II for conversational telephone speech for instance).
Secondly, as the CLIPS system, the LIA system was constrained to find two speakers only while processing conversational telephone data whereas it was unconstrained for the other kinds of data. This constraint on the number of speakers obviously improves the performance of the speaker segmentation system, when dealing with conversational telephone speech data.
Thirdly, the last evolution was made at the parameterization step for which Linear frequency Cepstral Coefficients (LFCC) were preferred to the MEL ones (MFCC), especially while processing broadcast news data, for which experiments showed performance improvement.

3.2.3. Common evolution

Prior acoustic macro-class segmentation
In 2003, the speaker evaluation campaign was focused on broadcast news data. As underlined in the previous section, various acoustic classes may be found in such data like studio speech, telephone speech, speech over music, music, ... Therefore, the CLIPS and LIA labs investigated the use of a prior macro-class acoustic segmentation before applying speaker segmentation systems, in order to:

- discard non speech signals (music, silence, ...), which have not to be processed by the speaker segmentation systems (otherwise, they involve speaker segmentation errors since they are labeled as speakers);
- provide a prior knowledge which may be interesting for the speaker segmentation systems, like gender detection or bandwidth classification (telephone vs non telephone speech).

Practically, the acoustic macro-class segmentation is first applied on each audio document, providing four sets of segments (bandwidth and gender detection), on which the speaker segmentation systems are applied individually. Finally, the segmentation outputs yielded on each individual acoustic macro-classes are merged to provide an overall segmentation.
The acoustic macro-class system used in this context relies on a hierarchical segmentation performed in three successive steps as illustrated in figure 1:

- during the first step, a speech / non speech segmentation of signal (representing a show) is performed using $MixS$ and NS models. The first model represents all the speech conditions while the second one represents the non speech conditions. Basically, the segmentation process relies on a frame-by-frame best model search. A set of morphological rules are then applied to aggregate frames and label 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 by 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, here, of three states (S, T, and SM models). The transition probabilities of this ergodic HMM are learnt on 1996 HUB 4 broadcast news corpus.
- the last step is devoted to gender detection. According to the label given 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 class ($GT-Fe$ and $GT-Ma$ for female and male telephone resp., $GS-Fe$ and $GS-Ma$ for clean speech, $GSM-Fe$ and $GSM-Ma$ for speech

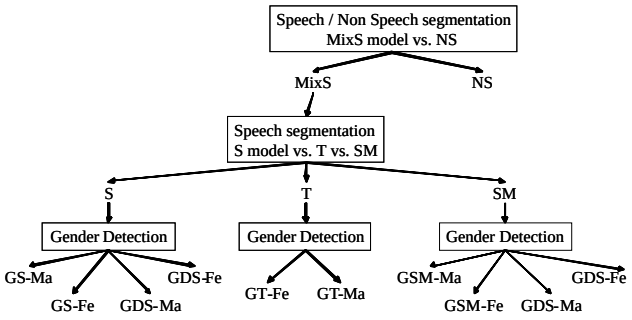

Fig. 1. *Hierarchical acoustic segmentation.*

over music, and $GDS - Fe$ and $GDS - Ma$, representing speech recorded over degraded conditions used to refine the final segmentation). The segmentation process, described in the previous step, is applied in the same way here.

All the state models mentioned above are diagonal GMMs except NS and $MixS$ models which are characterized by 1 and 512 Gaussian components respectively, all the other models are characterized by 1024 Gaussian components. They were trained on the 1996 HUB 4 broadcast news corpus.
Different levels of acoustic macro-class segmentation had been proposed and evaluated in terms of speaker diarization improvement. Results of this work may be found in [14].

Multi-channel meeting evaluation
In 2004 (Spring), the focus was made on meeting data and more precisely on multi-channel speaker segmentation. In this context, the speaker segmentation system has to process multiple speech channels coming from different microphones. Therefore, the choice of an efficient merging strategy in order to discard the irrelevant information becomes a crucial issue. This point will be discussed in the next section.
No particular tuning has been done on both the LIA and CLIPS speaker segmentation systems to participate at this evaluation campaign, except the use of a speech/non speech segmentation as a preliminary phase to deal with the specificities of meeting data (background noise, long silence, ...). This speech/non speech segmentation system consisted in a silence detection based only on a bi-gaussian modeling of the energy distribution associated with a detection threshold. The silence segment minimal length was set to 0.5s.

3.2.4. Strategies of fusion

In 2002, and more largely in 2003, the experience gained on the individual speaker segmentation system behavior (observed during the previous evaluation campaigns) leads the

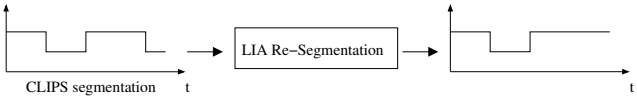

Fig. 2. *Example of the piped strategy.*

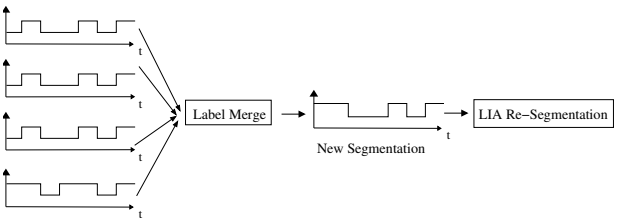

Fig. 3. *Example of the merging strategy.*

LIA and CLIPS labs to investigate various possibilities for combining their systems (for more details about these strategies, see [1, 2]).

Hybridization ("piped" system)
The hybridization strategy, illustrated in figure 2, consists in using the segmentation results of the CLIPS system to initialize a variant of the LIA system. Indeed, the speakers detected by the CLIPS system (number of speakers and associated audio segments) are used as an initialization of the LIA system HMM model (the models are trained using the information issued by the clustering phase), followed by an iterative process, during which adaptation and decoding steps are performed (similarly to the step 3 of the baseline LIA system). The process, involving a variant of the LIA system, is called the LIA re-segmentation process.
This solution associates the advantages of longer and (quite) pure segments of the step-by step strategy with the HMM modeling and decoding power of the integrated strategy.

Merging Strategy ("fusion" system)
The idea of "fusion" is to use the segmentations issued from as many experts as experts, as shown in figure 3. The merging strategy relies on a frame based decision which consists in grouping the labels proposed by each of the four systems at the frame level. Since two many virtual speakers are thus generated, the LIA re-segmentation process (described in the previous strategy), associated with some empirical rules, is applied to discard as many irrelevant speakers as possible.

Individual Microphone Segmentation Merging Strategy
A third strategy of fusion was especially designed in 2004 by the LIA and CLIPS labs to process multi-channel audio documents, relative to the meeting corpora. The goal of this strategy was to merge the multiple distant microphone seg-

mentations in a single meeting speaker segmentation output. Since no single signal is representative of the overall meeting, this strategy had to rely on some segment selection rules over the multiple distant microphone speaker segmentations. In this way, a specific merging algorithm was investigated. It relies on an iterative process which aims at detecting the longest speaker interventions over the set of distant microphone segmentations. Based on 3 successive steps, this iterative algorithm consists basically in selecting the longest speaker intervention over all microphone segmentation outputs taken separately, deleting in each distant microphone segmentation all the segments attributed to this new speaker, verifying the presence of not selected segments over all the distant microphone segmentations (more details on this algorithm are given in [3]).

4. CONTEXT AND PERFORMANCE COMPARISON

4.1. System evaluation

Table 1 presents a summary of the LIA and CLIPS (ELISA consortium) results obtained since 2000 on the different kinds of data: telephone speech conversations, broadcast news documents, and meeting room recordings (over various conditions: head microphone, or (distant) table microphone/mono-, or multi-channel signals).

Performance shown on the fifth line of the table illustrates the increasing difficulty of the tasks. Indeed, broadcast news and meeting data were introduced in the 2002 speaker segmentation system evaluation campaign. The ELISA systems were not really prepared for these new kinds of data. For telephone conversations, only two persons are involved, making the segmentation process easier. 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 first three lines illustrate the performance improvement of the LIA systems (LIA/EURECOM system for 2000) on telephone data over the three years. This performance was measured using the NIST 2000 metric (see the NIST evaluation plan for more details on this metric [9]), which was changed from the NIST 2002 evaluation campaign [11]. The third and fourth lines illustrate this change of the evaluation metric through the 2002 LIA system performance.
The sixth line shows the best performance obtained in 2003 on broadcast news data. This performance was achieved by the ELISA ”piped” system, which is based on the fusion strategy involving the LIA and CLIPS systems, as described in section 3.2.4. This line illustrates also the progress made from 2002 to 2003 on these data.
Finally, the last line presents the performance of the ELISA system during the NIST 2004 (Spring) speaker segmentation evaluation. Two different results are given, relating to the mono- or multi-channel speaker segmentation tasks. Despite the simplicity of the strategy used to merge the multiple channel segmentation outputs, the multi-channel speaker segmentation system obtained the best speaker diarization performance for the corresponding task [3].

4.2. Discussion

The participation of the LIA and CLIPS labs, every year since 2000, enforced by their strong collaboration, have enabled them to gain some experience on the speaker segmentation task and more precisely on the proposed approaches: step-by-step and integrated strategies.

First of all, the work done on the system fusion has highlighted the advantages and drawbacks of each individual approach. Indeed, through the piped strategy, it has been underlined the power of the E-HMM in modeling the conversation between speakers, which can be strongly enforced when the frontiers between speakers are provided by a robust BIC based speaker turn detection strategy.
Secondly, the evaluation of the E-HMM strategy over various kinds of data (conversational telephone speech, broadcast news, or meeting data) has shown the difficulty of controlling the E-HMM parameters, mainly based on heuristics. In the same way, another issue lying on the acoustic adaptation of the UBM model, used during the iterative adaptation/decoding process, was raised. Indeed, this acoustic adaptation currently relies on the speaker recognition algorithms, for which only mean parameters are adapted. In this context, it will be interesting to investigate more complex acoustic adaptation, involving either variance, weight, or both of them in order to take into account the difficulty of the task, compared with speaker recognition.

Moreover, it has been largely observed, over these different evaluation campaigns, that the characteristics of the signal files used to measure speaker segmentation system performance are very important to understand the behavior of the speaker segmentation system. The size of the signal file is the first factor. Indeed, the increase in terms of signal file duration (if implying an increase in the number of speakers), may induce some issues with the integrated approach, since the adaptation and decoding process may become less controllable, because of the lack of robustness of the current stop criterion. This last point remains an important issue of the Integrated approach, which will demand further investigation to find a more robust one. The clus-

tering phase of the step-by-step approach however seems to perform better as the file duration increases since the UBM used is trained directly from the entire test file. However, the drawback with this approach is the computational time which increases exponentially with the file duration. The variability of acoustic conditions observed in the files is also important. Indeed, the presence of large amount of telephone speech in the files may help the speaker segmentation system and therefore increases overall performance. Lastly, the number of signal files has to be sufficient in order to provide a robust evaluation measure. These different points have been largely discussed during the evaluation campaigns.

Speaker diarization on meeting data may be considered as a quite new task in the domain, particularly when the segmentation process has to deal with signals coming from multiple distant (table) microphones, scattered inside the meeting room. This multi-channel task has been introduced in 2004 in addition to the mono-channel one (meeting data used for the NIST-SpRec-2002 evaluation campaign being only mono-channel ie only based on one signal file to process, corresponding most often to the most informative microphone). Currently, the LIA and CLIPS systems, despite their proposed multi-channel fusion algorithm, do not really take advantage of the multi-channel information as shown by the small difference in terms of system performance between mono- and multi-channel (26.5% vs. 22.4% resp. in table 1). As underlined before, meeting data speaker segmentation remains, far from the other environments, the hardest one. The proposed speaker segmentation approaches cannot currently deal with a large part of issues, raised by this particular context: robust speech/non speech segmentation, robust speech vs background speech segmentation, voice overlapping inside one signal file but also between signal files, fusion of segmentation outputs without signal synchronization, etc. This long list shows that speaker segmentation process in the context of meeting data remains an open-domain.

Finally, in most research works on speaker segmentation, one of the main assumption is that there is no a priori information available on the test data. This means that there is no knowledge of the number or the identity of the speakers involved and in consequence there is no reference speaker data available for any one of them. This limitation may however not be so rough for certain applications and conditions where a priori data might be available. Generally the type of conversation is known (broadcast news, telephone or meeting). This gives us information on speech quality and average speaker turn length. In some cases, reference data might be available for some of the speakers. For broadcast news data for instance we can easily obtain reference data for the news host directly from the previous shows. A simple tracking system of this particular speaker can then lead to an error reduction. Some studies have been done in this sense by the authors (see [15]) and have shown the interest of a such approach. Further investigation has to be done in this way.

5. CONCLUSION

This paper presents the work done by the LIA and CLIPS labs since their first participation in the speaker segmentation system evaluation campaigns in 2001 (LIA only for this year). Over these four years, two main strategies have been proposed individually by each lab, to perform speaker diarization task: the CLIPS Step-by Step approach, based on a BIC detection criterion followed by a clustering process and the LIA integrated approach, based on a modeling of the speaker conversation (E-HMM).

The various range of data proposed during these evaluations: telephone conversational speech, broadcast news shows, as well as meeting data (mono-, or multi-channel signals), allows the LIA and CLIPS labs to evaluate their own strategy through different environments and to propose further evolution to improve each of them. Besides, the strong collaboration between the LIA and CLIPS labs allows to design some fusion strategy based systems, especially on the broadcast news data, involving both the Step-By-Step and Integrated approaches. This overall work, in addition to improve speaker segmentation system performance, allows to better understand the behavior of each individual approaches and to analyze the advantages and drawbacks of each of them. Particularly, it has been highlighted that the power of the Integrated approach should be enforced when robust frontiers between speakers can be provided by a robust BIC based speaker turn detection strategy (first step of the Step-By-Step approach). Further investigation will be done in this way, in order to really integrate both of them in a same system. Indeed, a future project launched by the LIA, CLIPS, and LIUM labs will be to develop a new speaker segmentation system, based on the free speaker recognition ALIZE toolkit [16], designed and developed in the framework of the ALIZE project, a part of the French research Ministry Technolangue program [17].

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] C. Fredouille, J.-F. Bonastre, T. Merlin, AMIRAL: a block-segmental multirecognizer architecture for automatic speaker recognition, Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3) (2000) 172–197.

[5] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: A review, The Annals of Statistics 6 (2) (1978) 461–464.

[6] 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.

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

[8] 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.

[9] NIST, The NIST year 2000 speaker recognition evaluation plan, `http://www.nist.gov/speech/tests/spk/2000/doc/spk-2000-plan-v1.0.htm` (January 2000).

[10] 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).

[11] 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).

[12] 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).

[13] 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).

[14] S. Meignier, D. Moraru, C. Fredouille, L. Besacier, J.-F. Bonastre, Benefits of prior acoustic segmentation for automatic speaker segmentation, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2004), Montreal, Canada, 2004.

[15] 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.

[16] Alize toolkit, `http://www.lia.univ-avignon.fr/heberges/ALIZE/`.

[17] French minister technolangue project, `http://www.technolangue.net/`.

Table 1. Best results (in %) achieved by the ELISA consortium along the different evaluation campaigns (since 2000), according to various corpora. *The three first line scores are computed according to the NIST 2000 metric, whereas the other results are computed with the RT metric (diarization speaker error rate)*

Corpus Evaluation Campaign	Telephone	Broadcast News	Meeting (head mic.) (mono-channel)	Meeting (table mic.) (mono-channel)	Meeting (table mic.) (multi-channel)
2000 LIA/EURECOM	31.0	X	X	X	X
2001 LIA	26.0	X	X	X	X
2002 LIA	10.0	X	X	X	X
2002 LIA	7.4	X	X	X	X
2002 LIA & CLIPS	5.7	30.3	34.7	36.9	X
2003 LIA & CLIPS	X	12.9	X	X	X
2004 LIA & CLIPS	X	X	X	26.5	22.4