

ELISA Nist RT03 Broadcast News Speaker Diarization Experiments

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

▶ **To cite this version:**

Daniel Moraru, Sylvain Meignier, Corinne Fredouille, Laurent Besacier, Jean-François Bonastre. ELISA Nist RT03 Broadcast News Speaker Diarization Experiments. The Speaker and Language Recognition Workshop (Odyssey 2004) , May 2004, Tolède, Spain. hal-01434300

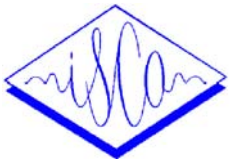
HAL Id: hal-01434300
https://hal.science/hal-01434300

Submitted on 22 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ISCA Archive
http://www.isca-speech.org/archive



ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

ELISA Nist RT03 Broadcast News Speaker Diarization Experiments

Daniel Moraru [(1)], Sylvain Meignier [(2)],
Corinne Fredouille [(2)], Laurent Besacier [(1)], Jean-François Bonastre [(2)]

[1] CLIPS-IMAG (UJF & CNRS) - BP 53 - 38041 Grenoble Cedex 9 - France
[2] LIA-Avignon - BP1228 - 84911 Avignon Cedex 9 – France

(daniel.moraru,laurent.besacier)@imag.fr
(sylvain.meignier,corinne.fredouille,jean-francois.bonastre)@lia.univ-avignon.fr

ABSTRACT

This paper presents the ELISA consortium activities in automatic speaker diarization (also known as speaker segmentation) during the NIST Rich Transcription (RT) 2003 evaluation. The experiments were achieved on real broadcast news data (HUB4), in the framework of the ELISA consortium. The paper firstly shows the interest of segmentation in acoustic macro classes (like gender or bandwidth) as a front-end processing for segmentation/diarization task. The impact of this prior acoustic segmentation is evaluated in terms of speaker diarization performance. Secondly, two different approaches from CLIPS and LIA laboratories are presented and different possibilities of combining them are investigated. The system submitted as ELISA primary obtained the second lower diarization error rate compared to the other RT03-participant primary systems. Another ELISA system submitted as secondary outperformed the best primary system (i.e. it obtained the lowest speaker diarization error rate).

1. INTRODUCTION

Speaker diarization (or segmentation) is a new speech processing task resulting from the increase in the number of multimedia documents that need to be properly archived and accessed. One key of indexing can be speaker identity. The goal of speaker diarization is to segment a N-speaker audio document in homogeneous parts containing the voice of only one speaker (also called speaker change detection process) and to associate the resulting segments by matching those belonging to a same speaker (clustering process). Generally, no *a priori* information is available on the number of speakers involved in the conversation as well as on the identity of the speakers.

The NIST Rich Transcription (RT) Evaluation[1] is sponsored in part of the DARPA Effective Affordable Reusable Speech To Text (EARS) Program. The EARS research effort is dedicated to developing powerful speech transcription technology that provides rich and accurate transcripts. It includes speech transcription but also acoustic segmentation, speaker indexing, disfluency detection induced by spontaneous speech (hesitations, self repairs, word fragments…), etc. Making available this rich transcription will authorize a better job when a machine is detecting, extracting, summarizing, and translating important information. EARS is focusing on natural, unconstrained human-human speech from broadcasts and foreign conversational speech in multiple languages.

This paper presents the ELISA Consortium [1] activities in automatic speaker segmentation during the NIST RT evaluation campaign organized in 2003.

The first part of this paper focuses on acoustic segmentation, which was mainly introduced to help automatic speech recognition (ASR) systems within the special context of broadcast news transcription. Indeed, at the beginning, one of the main objective of acoustic segmentation was to provide ASR system with an acoustic event classification allowing to discard non speech signal (silence, music, …), to adapt ASR acoustic models to some particular acoustic environments (like speech over music, telephone speech) or simply to speaker gender [2][3][4]. Many papers were dedicated to this particular issue and to the evaluation of acoustic segmentation in the context of ASR task. Nevertheless, rarely discussed in the literature, acoustic segmentation may be useful for other tasks linked to broadcast news corpora. In this sense, the aim of this paper is to investigate the impact of acoustic segmentation when applied as a prior segmentation for the particular task of speaker diarization.

The second part of the paper presents two systems – from CLIPS and LIA laboratories – which exhibit two different segmentation strategies. Various combination schemes of both systems are also investigated (the LIA macro class acoustic segmentation process is combined with the two speaker segmentation systems).

Section 2 presents the LIA macro class acoustic segmentation system, based on a hierarchical strategy. *Section 3* is dedicated to the presentation of the two speaker segmentation approaches involved in this work. Both begin by an acoustic pre-segmentation, also presented in this section. *Section 4* describes the combining strategies. The performance of the various propositions are shown and discussed in *Section 5* (the data are issued from RT 2003 evaluation campaign). Finally, *Section 6* concludes this work and gives some perspectives.

2. ACOUSTIC MACRO CLASS SEGMENTATION

Acoustic macro class segmentation is necessary to discard non speech signal (like music, silence, …) or to adapt acoustic models to specific acoustic environments (telephone speech,

[1] See http://www.nist.gov/speech/tests/rt/rt2003/index.htm for more details

speech over music, …). This is especially true when handling broadcast news data with the aim of automatically transcribing speech for instance. Basic speech/non speech detection is also useful for speaker segmentation task in order to avoid music portions to be automatically labeled as a new speaker. However, acoustic segmentation system may be designed to provide finer classification. For example, gender classification could help the segmentation process, by selecting the appropriate a priori knowledge. In this paper, the prior acoustic segmentation is done at three different levels: Speech/Non speech, Clean/Over music/Telephone[2] speech and Male/Female speech.

2.1 Front end processing

The signal is characterized by 39 acoustic features computed every 10 ms on 25 ms Hamming-windowed frames: 12 MFCC augmented by the normalized log-energy, followed by the delta and delta-delta coefficients.

2.2. Hierarchical approach

The system 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" 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 other models, "GDS-Fe" and "GDS-Ma", representing female and male speech recorded over degraded conditions are also used, at this stage, 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 [5] 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.

[2] This segmentation may also be referred to as a narrow/wide band speech classification if "speech over music" label is not used and simply considered as "speech" label.

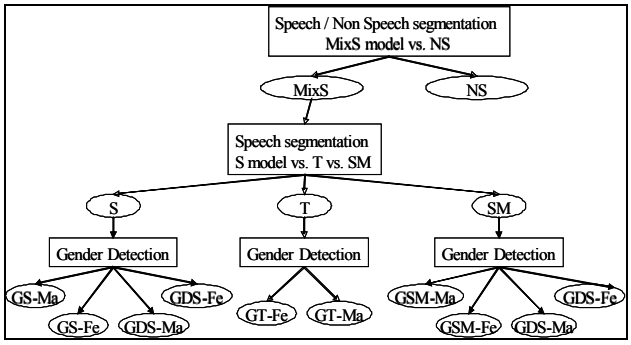


Figure 1: Hierarchical acoustic segmentation

3. SPEAKER SEGMENTATION SYSTEMS

All the speaker segmentation systems were developed in the framework of the ELISA consortium using AMIRAL, the LIA Speaker Recognition System [6].

Two different speaker segmentation systems are presented in this section. They have been developed individually by the CLIPS and LIA laboratories. Basically, the CLIPS system relies on a BIC-detector-based strategy followed by an hierarchical clustering [7]. The LIA system shows a different strategy, based on a HMM modeling of the conversation and an iterative process which adds the speakers one by one. Both of them use the acoustic pre-segmentation - described in section 2 - as a preliminary phase.

3.1. The LIA System

The LIA system is based on Hidden Markov Modeling (HMM) of the conversation [8][9]. Each state of the HMM characterizes a speaker and the transitions model the changes between speakers.

The speaker segmentation system is applied separately on each of the four acoustic classes detected by the acoustic segmentation described in section 2. Finally, the four segmentations are merged and a re-segmentation process is applied.

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 more 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 [8] or when no more speech is left to initialize a new speaker. A heuristic criterion is added to the likelihood-based criterion: if the last added speaker is tied to only one segment (<4sec), the previous segmentation is kept and a new speaker is added using the second best segment from *Step2*.

When the four (sub) segmentations are obtained independently using the previously described algorithm, they are merged and a re-segmentation phase starts.

The re-segmentation is similar to the adaptation and decoding step *(Step 3*). The main difference between the two phases is the GMM adaptation algorithm. The both adaptation algorithms are variants of MAP Bayesian adaptation. A classical MIT MAP adaptation [5] is used for the re-segmentation phase while a LIA variant optimized for an adaptation on a very short segment [6] is performed during the (sub)segmentation phase. In both adaptation phases only means are adapted.

The signal is characterized by 20 linear Cepstral features (LFCC) computed every 10 ms using a 20ms window. The Cepstral features are augmented by the energy (E). 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. The background models are trained on a subset of Broadcast News 96 data (F0, F1 and F2 acoustic conditions).

The LIA also presented a secondary system[3] closed to the previous one but using another variant of MAP for the speaker model adaptation. This algorithm is based on a linear combination of the estimated data and the a priori information. This adaptation method was employed by the LIA during NIST 2002 speaker recognition evaluation [10].

3.2. The CLIPS System

The CLIPS system [10] is based on a BIC (Bayesian Information Criterion) speaker change detector followed by an hierarchical clustering. The clustering stop condition is the estimation of the number of speakers using a penalized BIC criterion.

The speaker segmentation system is applied separately on each of the four acoustic classes detected by the acoustic segmentation described in section 2. A BIC [7] approach is then used to define first potential speaker changes. 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 by first training a 32 components GMM background model (with diagonal covariance matrices) on the entire test file maximizing a ML criterion thanks to a classical EM algorithm. Segments 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).

The number of speakers in the conversation (NSp) is estimated using a penalized BIC (Bayesian Information Criterion), in contrast with the last year CLIPS segmentation system which used a fixed number of speakers [10].

The number of speakers is constrained between 1 (if we are working on an isolated acoustic pre-segmentation class) 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 (NSp) is selected to maximize:

$$BIC(M)=\log L(X;M)-\lambda\frac{m}{2}NSp\log NX$$

where M is the model composed of the NSp speaker models, NX 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 signal is characterized by 16 mel Cepstral features (MFCC) computed every 10ms on 20ms windows using 56 filter banks. Then the Cepstral features are augmented by energy. No frame removal or any coefficient normalization is applied.

4. COMBINING STRATEGIES

In this section we investigate two possibilities for combining the systems, firstly using an hybridization strategy and secondly by merging the proposed segmentations. The merging strategy is a new way of combining results coming from multiple segmentation systems.

4.1 Hybridization ("piped" system)

The idea of this hybridization strategy is to use the results of the CLIPS system to initialize the LIA re-segmentation system (figure 2). The speakers detected by the CLIPS system (number of speakers and associated audio segments) are inserted in the re-segmentation HMM model (the models are trained using the information issued by the clustering phase). This solution associates the advantages of longer and (quite) pure segments with the HMM modeling and decoding power.

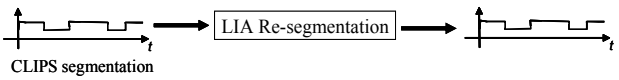


Figure 2: ELISA piped system

4.2 Merging Strategy ("fusion" system)

The idea of "fusion" is to use the segmentations issued from as many as possible experts, four in this paper (figure 3): CLIPS primary system, LIA primary system, LIA secondary system, ELISA piped system.

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. An example (for four systems) is illustrated below:

- Frame *i*: Sys1="S1", Sys2="T4", Sys3="S1", Sys4="F1" → label result "S1T4S1F1",
- Frame *i+1* : Sys1="S2", Sys2="T4", Sys3="S1", Sys4="F1" → label result "S2T4S1F1".

[3] This system is used in section 3.2 ("fusion" system).

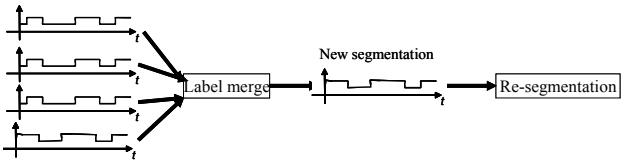


Figure 3: ELISA merge system

This label merging method generates (before re-segmentation) a large set of virtual speakers (~150 virtual speakers per show) composed of:

- Virtual speakers that have a large amount of data assigned. These speakers could be considered as correct hypothesis speakers;
- Virtual speakers generated by few systems, for example the speakers associated with only one short segment (~3s up to 10s). These hypothesis speakers could be suppressed (the weight of these speakers on the final scoring is marginal);
- Virtual speakers that have a smaller amount of data scattered between multiple small segments and that could be considered as zones of indecision.

Based on these considerations, the LIA re-segmentation is then applied on the merged segmentation. Between each adaptation / decoding phase, the virtual speakers for whom total time is shorter than 3s are deleted. The data of these deleted speakers will further be dispatched between the remaining speakers during the next adaptation / decoding phase.

After the first iteration 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 could split a "true" speaker in two hypothesis speakers, each tied to a long segment. Suppose all systems agreed on a long segment except one which splits it in two parts. This would produce two virtual speakers (associated with long duration segments) after the merging phase and since we are not doing any clustering before re-segmentation, we would have a "true" speaker spitted in two virtual speakers.

5. EXPERIMENTS AND RESULTS

The experiments described in this paper were conducted in the framework of NIST/RT'03[4] evaluation campaign. In this context, two separate corpora were available:

- *Dev* corpus composed of 6 broadcast news shows of 10mn each, recorded in 1998, available to tune the speaker segmentation systems;
- *Eva* corpus composed of 3 broadcast news shows of 30mn each, recorded in 2001, each containing between 10 and 27 speakers and available for evaluation only.

5.1 Speaker Diarization performance measure

In order to measure the accuracy of the segmentation, we used the scoring system proposed during RT'03 evaluation, a scoring based on the percentage of frame well classified.

Firstly, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs is computed. The measure of optimality is the aggregation, over all reference speakers, of time that is jointly attributed to both the reference speaker and the (corresponding) system output speaker to which that reference speaker is mapped. The resulting time-based speaker diarization error score corresponds to the fraction of speaker time that is not attributed correctly to a speaker. This scoring takes into account both speaker segmentation error (speech segments attributed to a wrong speaker) and missed/false alarm speaker errors, directly linked to speech/non speech classification errors.

5.2. Raw performance of the prior acoustic segmentation

Table 1 gives the performance of the prior acoustic segmentation process, taken alone. The speech/non speech segmentation error is around 4.5% (in terms of duration) compared to 4.4% for the best system during NIST RT'03 evaluation campaign. The gender detection error goes from 1.5% for the *Dev* set to 5.5% for the *Eva* set. Thanks to the manual segmentation Hand S/NS-Gender-T/NT, the accuracy of acoustic segmentation system at the level of telephone and non telephone classification is evaluated: less than 0.1% for *Dev* corpus and 3% for *Eva*.

Corpus	Missed Speech Error	False Alarm Speech Error	Gender Error	Telephone / Non telephone Speech error
Dev	**2.3%**	2.2%	1.5%	0.09 %
Eva	1.8%	2.7%	5.5%	3 %

Table 1: Acoustic segmentation errors on Dev and Eva sets.

5.3. Impact of acoustic macro class segmentation on speaker diarization performance

For evaluating the impact of acoustic macro class segmentation on speaker diarization performance, different levels of acoustic segmentation granularity are evaluated on both speaker segmentation systems:

- Speech/non speech classification only (S/NS);
- segmentation based on speech/non speech and gender detection (S/NS-Gender);
- segmentation based on speech/non speech, gender and telephone/non telephone speech detection (S/NS-Gender-T/NT);
- segmentation based on speech/non speech, gender and telephone/clean speech/speech over music/degraded speech (S/NS-Gender-T/S/MS/DS).

For comparison, some speaker segmentation results will be also presented based on an acoustic segmentation, obtained manually and based on speech/non speech, gender and telephone / non telephone speech detection (Hand S/NS-Gender-T/NT).

For both corpora, the advertisement portions were manually discarded before any treatment which explains that results presented in this section do not correspond exactly to official RT'03 results and to ones presented in the next section.

Table 2 and table 3 present the experimental results obtained when combining speaker segmentation systems (speaker turn detection based and ascending/HMM based approaches) with acoustic segmentation before and after the re-segmentation step respectively. These results, expressed in terms of diarization error rates, show :

[4] More information about RT03 Evaluation could be find at: http://www.nist.gov/speech/tests/rt/index.htm

- a large variation in terms of performance between *Dev* and *Eva* corpora, especially for the ascending/HMM approach;
- a gain in performance (the best one) for both speaker segmentation systems when they are combined with manual acoustic segmentation;
- a significant improvement of ascending/HMM approach results on *Eva* corpus due to speech/non speech, gender and telephone/non telephone segmentations (S/NS-Gender-T/NT) without (from 26.9% to 18.1%) and with (from 26.6% to 14.1%) re-segmentation phase. On *Dev* corpus, this improvement is visible only after re-segmentation phase (from 15.5% to 12.8%);
- no real effect on speaker-turn-based approach regarding the different levels of acoustic segmentation granularity, except on *Eva* corpus after re-segmentation step for which result improvement can be observed (from 15.7 to 13.9%);
- no improvement or even performance loss when the finest acoustic segmentation (S/NS-Gender-T/S/MS/DS) was involved;
- the interest of the re-segmentation step for both speaker segmentation strategies since best performance may be observed in most of the cases.

	Ascend./HMM approach		Speaker turn approach	
Acoustic segmentation	***Dev***	***Eva***	***Dev***	***Eva***
Hand S/NS-Gender-T/NT	*13.1%*	*15.4%*	*14.1%*	*10.6%*
S/NS	15.4%	26.9%	19.7%	17.1%
S/NS-Gender	14.8%	27.4%	18.7%	18.6%
S/NS-Gender-T/NT	15.1%	18.1%	19.0%	18.2%
S/NS-Gender-T/S/MS/DS	25.9%	30.5%	19.1%	27.6%

Table 2: Diarization error rates for different levels of acoustic granularity on Dev and Eva sets before re-segmentation step.

	Ascend./HMM approach		Speaker turn approach	
Acoustic segmentation	***Dev***	***Eva***	***Dev***	***Eva***
Hand S/NS-Gender-T/NT	*10.8%*	*13.2%*	*15.7%*	*9.4%*
S/NS	15.5%	26.6%	19.3%	15.7%
S/NS-Gender	13%	24.9%	18.2%	15.3%
S/NS-Gender-T/NT	12.8%	14.1%	19.0%	13.9%
S/NS-Gender-T/S/MS/DS	15.6%	14.3%	18.7%	15.1%

Table 3: Diarization error rates for different levels of acoustic granularity on Dev and Eva sets after re-segmentation step.

	Miss Speech	FA Speech	SPK ERR	**ERR**
CLIPS primary	2.0%	2.9%	14.3%	**19.25%**
LIA primary	1.1%	3.8%	12.0%	**16.90%**
LIA second	1.1%	3.8%	19.8%	**24.71%**
ELISA "merged"	1.1%	3.8%	9.3%	**14.24%**
ELISA "piped"	1.1%	3.8%	8.0%	**12.88%**

Table 4 Experimental results on RT 2003 data

5.4. Performance of ELISA systems during RT03 evaluation

The fusion system submitted as ELISA primary system obtained the second lower segmentation error rate compared to the other RT03-participant primary systems. The ELISA pipe system submitted as secondary system outperformed the best primary system and obtained the lowest speaker segmentation error rate.

The table 4 summarizes the performance achieved by the different proposed systems during RT03. It shows that:

- Even if the five systems are based on the same acoustic segmentation, the *Miss Speech* and *False Alarm Speech* errors are different. This is due to the LIA and ELISA system behavior which work at 0.2s block level (all the segments boundaries are aligned on a 0.2s scale) whereas CLIPS system works at a frame level. It gives small differences in the border positions of the segments but the sum (Miss+FA) remains the same.
- The LIA and CLIPS systems obtained satisfactory results, compared to the other RT03 participants. The LIA HMM based primary system outperforms slightly the CLIPS classical approach (16,9% of total error compared to 19,25%). But the second LIA system - with a different model adaptation strategy - obtained only 24,71% of total error This result illustrates the difficulty of adapting a large statistical model in borderline conditions (only few seconds of adaptation data).
- The "piped" technique improves the performance. Giving good segment boundaries to the HMM based method increases drastically the performance (from 16,9% to 12,88% of total error). Indeed the re-segmentation phase improves the accuracy of the CLIPS segmentation and allows to reduce the segmentation error by 33% (relative).
- The "merged" strategy performs better than the "piped" strategy over two recordings (8% relative gain). Unfortunately a drastic loss is observed on the last recording. The lose on that particular recording is a good example of the limitation of the merging technique explained in 4.2: one of the systems disagreed with the others. This resulted in too many speakers detected and, most important, in a long speaker split in two that generated an important error[5].

For the CLIPS system, complementary experiments showed that estimating automatically the number of speakers during the clustering process generates only about 3% more of absolute segmentation error than the optimal number of speakers[6]. The CLIPS algorithm missed only 7% of the real speakers involved in the files (4 speakers out of 57 total speakers).

[5] The problem could also come from the nature of the test file: it is the only one narrow band file.

[6] The optimal number of speakers is the number of speakers that minimizes the segmentation error and not the real number of speakers involved in dialogue. Usually the optimal number is smaller than the real number. This is due to the fact that in the conversations some speakers pronounced only very short utterances and missing them does not have a significant effect on the total diarization error rate.

6. CONCLUSIONS

This paper summarizes the ELISA Consortium strategies for the speaker segmentation task. The ELISA effort was focused in the framework of NIST 2003 speaker diarization evaluation campaign.

This paper firstly investigated the impact of *prior* acoustic macro class segmentation when it is combined with speaker segmentation. This impact was evaluated in terms of speaker diarization error rate, according to NIST/RT'03 scoring. The *prior* acoustic macro class segmentation, presented in this paper, was developed by the LIA lab. It relies on a GMM model based hierarchical segmentation, designed to provide different levels of segmentation granularity (from simple speech/non speech detection to gender dependent acoustic classes such as speech over music, degraded speech or telephone speech). Experiments based on the combination of each speaker segmentation system with the acoustic segmentation were conducted according to the different levels of acoustic segmentation granularity. The results presented in this paper show the benefit of acoustic segmentation for speaker segmentation performance, especially for the ascending/HMM approach, while the speaker turn detection based system seems to be less dependent of an acoustic segmentation.

Then, we described two approaches for speaker diarization, the LIA system, based on a HMM modeling of each conversation (where all the information is reevaluated at each detection of a new speaker or a new segment), and the CLIPS system, which uses a standard approach based on speaker turn detection and clustering. Despite the differences between the approaches, the results obtained during the NIST RT03 evaluation showed the interest of each technique.

Several ways of combining the two systems were also proposed. The "piped" system improved significantly the performance, up to 33% of relative error reduction (from 19.25 % to 12.88%) and achieved the best performance during RT03 evaluation. A complete analysis of the results is necessary, to understand which part of the gain comes from the various ways of processing the information and which part comes from the correction of each system intrinsic errors.

One of the main drawback of both systems is the difficulty to detect the minority speakers that do not speak very much. Depending of the nature of the audio files, they could generate a large part of the segmentation errors.

As a perspective, we are currently working on adding to the conversation model a priori information for the segmentation system [11], for both ELISA approaches. Further work should also study the way of taking benefit of finer acoustic classes such as speech over music or degraded speech for the speaker segmentation task.

REFERENCES

[1] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium, "Overview of the 2000-2001 ELISA consortium research activities," *A Speaker Odyssey*, pp.67–72, Chania, Crete, June 2001.

[2] P.C. Woodland, "The development of the HTK Broadcast News transcription system: An overview", *Speech Communication*, Vol. 37, pp. 291-299, 2002.

[3] T. Hain, and P.C. Woodland, "Segmentation and Classification of Broadcast News audio", *ICSLP'98*, Sydney, Australia.

[4] J.L. Gauvain, L. Lamel, and G. Adda. "The LIMSI Broadcast News Transcription System". *Speech Communication*, 37(1-2):89-108, 2002.

[5] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adaptation Mixture Models". *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[6] C. Fredouille, J.-F. Bonastre, and T. Merlin, "AMIRAL: a block-segmental multi-recognizer architecture for automatic speaker recognition," Digital Signal Processing, Vol. 10, No. 1-3, January/April/July 2000.

[7] P. Delacourt and C. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing," *Speech Communication,* Vol. 32, No. 1-2, September 2000.

[8] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," *A Speaker Odyssey*, pp.175-180, Chania, Crete, June 2001.

[9] S. Meignier, J.-F. Bonastre, C. Fredouille, and T. Merlin, "Evolutive HMM for Multi-Speaker Tracking System". *ICASSP'00*, 5-9 June 2000, Istanbul, Turkey.

[10] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation". *ICASSP'03*, Hong Kong.

[11] D. Moraru, L. Besacier, E. Castelli, « Using a priori information for speaker diarization », Other Paper Submitted to Odyssey 2004.