THE ELISA CONSORTIUM APPROACHES IN SPEAKER SEGMENTATION DURING THE NIST 2002 SPEAKER RECOGNITION EVALUATION

Daniel Moraru [1], Sylvain Meignier [2]
Laurent Besacier [1], Jean-François Bonastre [2], Ivan Magrin-Chagnolleau [3]

[1] CLIPS-IMAG (UJF & CNRS) - BP 53 - 38041 Grenoble Cedex 9 - France
[2] LIA-Avignon – BP1228 – 84911 Avignon Cedex 9 - France
[3] Laboratoire Dynamique Du Langage (CNRS & University of Lyon 2) – 14, avenue Berthelot – 69363 Lyon Cedex 07 – France

daniel.moraru@imag.fr - sylvain.meignier@lia.univ-avignon.fr
laurent.besacier@imag.fr - jean-francois.bonastre@lia.univ-avignon.fr - ivan@ieee.org

ABSTRACT

This paper presents the ELISA consortium activities in automatic speaker segmentation during last NIST 2002 evaluation: two different approaches from CLIPS and LIA laboratories are presented and the possibility of combining them either by applying them consecutively, or by fusing the decisions made by each of them, is investigated. Various types of data were available for NIST 2002. The ELISA systems obtained the lower error rates for two corpora: the CLIPS system obtained the best performance on the Meeting data, the LIA system obtained the best performance on the Switchboard data. The combining strategies proposed in this paper allowed us to improve the performance of the best single system on both data types (up to 30 % of error rate reduction).

1. INTRODUCTION

Speaker indexing is a new task linked to speech processing 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. More precisely, from an algorithmic point of view, three different tasks can be pointed out in this domain. *Speaker tracking* consists in finding, in an audio document, all the occurrences of a particular speaker. This requires that this speaker is known *a priori* by the system (i.e. a model of his/her voice is available). In that sense, *speaker tracking* can be seen as a speaker verification task applied locally along a document containing multiple (and unknown) interventions of various speakers. The begin/end points of the tracked speaker interventions have to be found during the process. On the other hand, the goal of *speaker segmentation* – the task addressed in this paper – is to segment a N-speakers conversation 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 and identity of speakers involved in the conversation. Finally, *speaker tying* is a classification process consisting in finding the number of speakers present in a collection of audio documents segmented independently (speaker segmentation task) and to attribute the various utterances to the corresponding speaker [5].

This paper presents the ELISA Consortium [3] activities in automatic speaker segmentation during the NIST automatic speaker recognition evaluation campaign organized in 2002 (http://www.nist.gov/speech/tests/spk/). Two systems – from CLIPS and LIA laboratories – are presented and various combination schemes of both systems are investigated.

Section 2 is dedicated to the presentation of the two speaker segmentation approaches, while *Section 3* describes the proposed combining strategies. The performance of the various propositions are shown and discussed in *Section 4* (All the experimental protocols and data are issued from NIST 2002 evaluation campaign). Finally, *Section 5* concludes this work and gives some perspectives.

2. SPEAKER SEGMENTATION SYSTEMS

All the speaker segmentation systems were developed in the framework of the ELISA consortium [3]. The 2002 ELISA platform is based on AMIRAL, the LIA speaker recognition system [2]. Systems presented here are primary systems of the ELISA sites which competed during last NIST speaker verification / segmentation evaluations in spring 2002.

2.1. LIA Primary System

The LIA primary system is based on a hidden Markov modeling (HMM) of the conversation [6][4]. Each state of the HMM characterizes a speaker and the transitions model the changes between speakers (*Figure 1*).

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 S0 is trained on the whole test utterance. The segmentation is modeled by a one-state HMM and the whole signal is set to speaker S0.
Step 2-Adding a new speaker. A new speaker model is trained using the 3s of test that maximize the sum of likelihood ratios for model S0. A corresponding state, labeled Sx (x is the number of the iteration), is added to the previous HMM.
Step 3-Adapting speaker models. First, all the speaker models are adapted according to the current segmentation. Then, Viterbi decoding 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 stop criterion is based on the comparison of the probability along the Viterbi path between two iterations of the process [4] and on the number of segments labeled Sx (If the last added speaker Sx is tied to only one segment, the previous segmentation is kept and we stop).
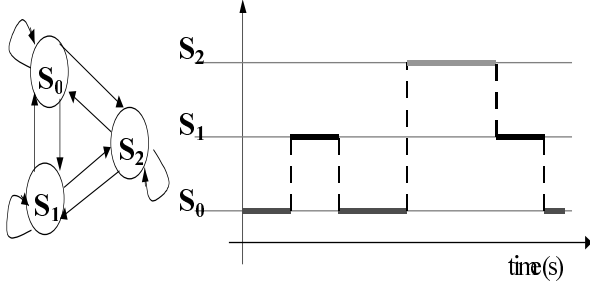

Figure 1: LIA/HMM modeling of the conversation.

The signal is characterized by 20th order linear cepstral features (LFCC) computed at a 10 ms frame rate using a 20ms window. Then the cepstral features are augmented by the energy (E). No frame removal or any coefficient normalization is applied. Speaker models are derived from a background model by MAP adaptation (means only). GMM with 128 components (diagonal covariance matrix) are used. The background models are trained on Switchboard II phase II data The HMM emission probabilities – for each 0.3s of the input stream and each HMM state – are estimated by computing the mean log likelihood ratio between the corresponding speaker model, background model and input segment.

2.2. CLIPS Primary System

2.2.1. Speaker Change Detection

Speech activity detection (SAD) is first applied on the signal. The SAD marks are used to define first potential speaker changes. A Bayesian Information Criterion (BIC: for more details see [1]) approach is then used. A BIC curve is extracted from 1.75s adjacent windows. Mono-Gaussian models with diagonal covariance matrices are used to build the BIC curve and the parameters are 16 MFCC+Energy coefficients with no Cepstral Mean Subtraction (CMS). A threshold is then applied on the BIC curve to find speaker changes. The threshold is tuned so that over-segmentation (more speaker changes detected) is provided since we prefer to detect more segments (which can be further merged by the clustering process) than missing speaker changes which will never be recovered later.

Another system was presented to the NIST 2002 evaluation with *a priori* segmentation using fixed length segments (0.75s). It gave approximately the same performance while being 3 times slower due to the uniform segmentation that leads to much more segments at the entry of the clustering module.

2.2.2. Clustering

First, a diagonal 32 GMM background model is trained on the entire file. Segments models are then trained using MAP adaptation (means only). BIC 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 in the conversation). In the *primary system*, N was

always set to 2 whatever the type of data was. However, as explained in the next section, the number N of speakers found for each test signal needs to be the same for both CLIPS and LIA systems before combination. Thus, we also built a *secondary system* for which N was the same as the N found by the LIA system.

Re-segmentation is then performed after clustering by building N speaker models from the segmented file. Likelihood scores are computed on 0.8 second segments to decide to which speaker Li (1<i<N) the segment belongs.

2.3. Main Differences Between Both Approaches

Table 1 summarizes the main differences between the LIA and CLIPS approaches.

System	Parameters	Segmentation	Clustering	Re-seg.
LIA	20LFCC+E	*a priori*	Descendant	N-A[1]
	No CMS	0.3s segments[2]	Estimate N	
CLIPS	16MFCC+E	BIC	Ascendant	yes
	No CMS		N fixed *a priori*	

Table 1: Overview of LIA and CLIPS systems.

3. COMBINING STRATEGIES

We investigated two directions for combining our systems, firstly using a hybridization strategy and secondly by fusing the proposed segmentations.

3.1. Hybridization

The idea of hybridization is to use the results of one system to initialize the other one; the segments found by the first system give first speaker change points for the second system. We experimented both possible configurations: LIA segmentation piped in CLIPS system and CLIPS segmentation piped in LIA system (*Figure 2*)
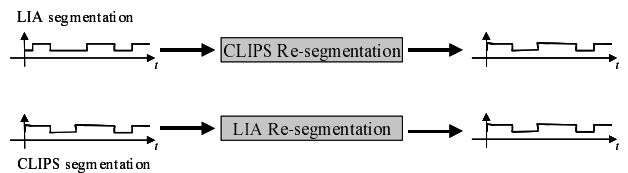

Figure 2: Hybridization of systems

3.2. Fusion

The idea of fusion is to use both segmentations of the two experts and to match the speaker segments as the NIST speaker segmentation scoring program does between the reference

---

[1] The LIA method is based on an iterative process which re-evaluates all the decisions at each iteration.
[2] The LIA method does not need any *a priori* segmentation but a segmentation in 0.3 s segments is done in order to save computation time.

segmentation and a hypothesized segmentation. The difference is that, in this case, there is no reference but a segmentation hypothesized by a second system. We suppose that both systems have found the same number of speakers in the conversation; so for fusion, the *secondary* CLIPS system is used (clustering with the value of N fixed by the LIA system, as explained in *Section 2.2.2*). The common segments (on which both experts agree) are kept while for the other segments, a new re-segmentation is done, by one system or another (CLIPS or LIA). The LIA re-segmentation is based on the "adaptation/decoding" step of the LIA segmentation system (*cf. 2.1 step 3*). In this case, the re-segmentation is initialized according to an initial segmentation given by the CLIPS.

The interest of this approach is that now we have an idea of the segments in which we can "trust" and only these common segments will be used to build the N speaker models and make a re-segmentation of the whole conversation. *Figure 3* shows the general principle of fusion of systems for segmentation.
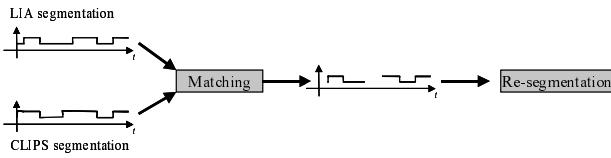

Figure 3: Fusion of systems.

4. EXPERIMENTS AND RESULTS

4.1. The NIST 2002 Speaker Segmentation Evaluation

Various types of conversations were given for the NIST 2002 speaker segmentation evaluation:

- 199 test segments (two minutes each) taken from Switchboard Cellular Phase 2 (SB) and involving only two speakers (8khz data);
- 83 test segments (two minutes each) taken from NIST recorded meetings (ME) involving various numbers of persons (N=4 to 6). Two versions of each segment (83 + 83 = 166 total) were available since meeting were simultaneously recorded with head mounted microphones and with table mounted microphone (16khz data);
- 76 test segments of broadcast news (BN), of variable length (35 – 142 seconds), taken from various Hub-4 corpora; involving various number of persons (mostly N=2 to 7, 16 kHz data).

The performance measure used for the NIST 2002 speaker segmentation task is the segmentation cost function, defined as a weighted sum of decision errors, weighted by error type and integrated over error duration. For speaker segmentation, there are five kinds of errors that can occur, all as a function of time:

- Missing a segment of speech when speech is present (PMissSeg)
- Falsely declaring a segment of speech when there is no speech (PFASeg)
- Assigning a false alarm speaker to a segment of speech (PMissSpkr)
- Assigning a speaker to a segment of speech of a missed speaker (PFASpkr)

- Assigning an incorrect speaker to a segment of speech (PErrSpkr)

The speaker segmentation cost is therefore defined as:

$$C_{Seg} = \left(C_{MissSeg} \cdot P_{MissSeg} + C_{FASeg} \cdot P_{FASeg} \right) + \left(C_{MissSpkr} \cdot P_{MissSpkr} + C_{FASpkr} \cdot P_{FASpkr} \right) + C_{ErrSpkr} \cdot P_{ErrSpkr}$$

The cost parameters are all set equal to 1.

Since there is no predefined speaker set, the set of speakers that the speaker segmentation system defines must be matched with the set of speakers that the answer key contains in order to minimize the cost function.

In the results presented further in this paper, this *Cseg* score is used to evaluate performance; the areas with overlapping speech (two speakers speaking at the same time) are also ignored during the scoring.

4.2. NIST2002 Evaluation Results

The results obtained during the NIST evaluation are given in *Table 2* for the systems alone, and then for hybridization and fusion. CLIPS primary system was not combined with the LIA primary system because it makes the hypothesis that N=2 speakers are involved in the conversation; therefore, for combination purpose, CLIPS secondary system was used (N fixed by LIA; same results between primary and secondary observed on SB data since exactly N=2 speakers are involved). All these results can be found on the *NIST 2002 Speaker Recognition Evaluation CD-ROM* distributed by NIST.

The *Baseline* indicates the "difficulty" of the task, since it is the score given by a system that basically decides that the entire test signal was uttered by a single speaker.

	BN	ME	SB
Baseline	48.4%	50.1%	30.8%
CLIPS primary (N=2)	**30.3%**	**35.8%**	8.6%
CLIPS secondary (used for combination)	34.2%	36.4%	8.6%
LIA primary (used for combination)	38.2%	40.2%	**7.4%**
Hybridization of systems			
LIA results followed by CLIPS re-segmentation **H1**	38.4%	39.2%	7.0%
CLIPS results followed by LIA segmentation **H2**	34.3%	36.3%	**6.0%**
Fusion of systems			
Fusion + CLIPS re-segmentation **F1**	33.7%	37.0%	7.6%
Fusion + LIA segmentation **F2**	33.6%	**35.0%**	**5.7%**

Table 2: Experimental results on NIST 2002 data.

The results show that:

- All the combining techniques (hybridization or fusion) improve the performance for SB corpus. It seems that the better the experts are, the better the combination is.
- Fusion of systems leads to the best performance for SB and ME corpora, in particular fusion followed by LIA segmentation.
- Fusion of two single systems improves their performance on BN data but performs worse than CLIPS primary system (number of speaker fixed to 2).

Looking at the performance separately on each of the 199 Switchboard conversations, we noticed that fusion systems F1 and F2 improved the performance on respectively 51 % and

70 % of the files compared to the best system. On the remaining files, fusion degraded the performance either because the results of a single system were already very good and difficult to improve, or because there was not enough matching between both systems decisions (high *Cseg* score between both systems), which led to an insufficient amount of data for building the re-segmentation models. To conclude, the fusion should not be used when not enough speech material is available for building re-segmentation models, namely when the two systems do not agree on enough segments. For this, a threshold on the *Cseg* score calculated between the decisions of the two systems (Cseg between LIA and CLIPS was evaluated to 14 % on average on SB data) could be applied; if this score is too high, one can then decide to cancel fusion for this conversation.
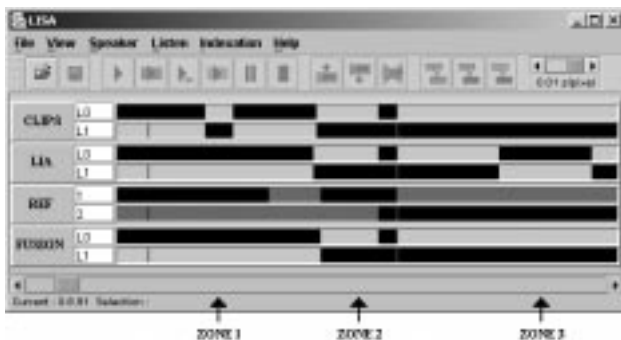

Figure 4: Example of fusion.

In order to show the positive effect of fusion, *Figure 4* presents the results on one part of a file. We can see that both systems can correct each other errors (ZONE1: LIA corrects CLIPS errors; ZONE2: both systems are wrong and nothing could be done; ZONE3: CLIPS corrects LIA errors).

4.3. Potentiality of Decision Fusion

Finally, we also calculated the score corresponding to the best "decision-based" theoretical fusion of LIA and CLIPS systems on Switchboard data. This is achieved by keeping the decision of these systems when they agree, and by taking the correct decision when they do not agree (on Switchboard, there are only two speakers, so when both systems do not agree, one of them is necessarily right). In other words, that would be the best fusion achieved if we were able to find a fusion strategy which takes the best possible decision on segments where the two systems disagree. This score is 2.9 %. This is the asymptotic fusion score given the LIA and CLIPS systems. It means that there is still a margin for progress in the fusion strategy itself.

5. CONCLUSION

This paper summarizes the ELISA Consortium strategies for the speaker segmentation task. We described 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, clustering and re-segmentation. Despite the differences between the approaches, the results obtained during the NIST 2002 evaluation showed the interest of each technique: the two systems obtained the best

results, respectively for Switchboard and Meeting data. The results were less encouraging on the BN data.

Several ways of combining the two systems were also proposed. The fusion of the two experts improved significantly the performance, up to 30 % of error reduction (from 7.4 % of error for Switchboard – best performance during NIST 2002 – to 5.7 %). 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 the system intrinsic errors. As a guideline, we calculated an asymptotic value for the best "decision-based" possible fusion of 2.9 % on Switchboard.

The main drawback remains the detection of the number of speakers involved in the conversation, since LIA system overestimates the number of speakers and CLIPS system fix it *a priori*. A better modeling of the conversation (duration models) is also an interesting way to improve the results, especially with the LIA HMM-based system. Finally, adding the detection of other meta-information (gender and channel) will certainly improve the results and we are currently working on these improvements.

6. REFERENCES

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

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

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

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

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

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