THE ELISA CONSORTIUM APPROACHES IN BROADCAST NEWS SPEAKER SEGMENTATION DURING THE NIST 2003 RICH TRANSCRIPTION EVALUATION

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 segmentation also known as speaker diarization during the NIST Rich Transcription (RT) 2003 evaluation. The experiments were conducted on real broadcast news data (HUB4). Two different approaches from CLIPS and LIA laboratories are presented and different possibilities of combining them are investigated, in the framework of ELISA consortium. The system submitted as ELISA primary system obtained the second lower segmentation error rate compared to the other RT03-participant primary systems. Another ELISA system submitted as secondary system outperformed the best primary system and obtained the lowest speaker segmentation 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. 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 involved in this work. Both begin by an acoustic pre-segmentation, also presented in this section. *Section 3* describes the combining strategies. The performance of the various propositions are shown and discussed in *Section 4* (All the experimental protocols and data are issued from RT 2003 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 using AMIRAL, the LIA Speaker Recognition System [2].

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 [3]. 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 below - as a preliminary phase.

2.1 Prior acoustic macro-class segmentation

A prior acoustic segmentation is necessary in order to suppress the non-speech segments. The non-speech segments decrease the accuracy of the segmentation process (and the speech/non-speech classification errors - missed speech error and false-alarm speech error - are penalized by the evaluation metric). A finer acoustic pre-segmentation is also useful in order to optimize the clustering process. For example, a gender pre-segmentation suppresses cross-gender cluster building and authorizes gender specific a priori knowledge (gender dependent world model for example).

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

The acoustic pre-segmentation used during RT03 evaluation provides a speech classification in:

- Speech / Non-speech
- Studio / Telephone (wide / narrow band) speech
- Male / Female speech.

It is important to note that the individual (CLIPS or LIA) segmentation strategies are applied on each (speech) separate acoustic conditions (Male Wide, Male Narrow, Female Wide, Female Narrow) yielded by this pre-segmentation.

The acoustic pre-segmentation system relies on a hierarchical scheme, which permits to refine the segmentation outputs at each level. It involves classical techniques like Cepstral parameterization, Gaussian Mixture Modeling and Viterbi decoding. A more precise description of this system may be found in [4]

2.2. The LIA System

The LIA system is based on Hidden Markov Modeling (HMM) of the conversation [5][6]. Each state of the HMM characterizes a speaker and the transitions model the changes between speakers (figure 1).

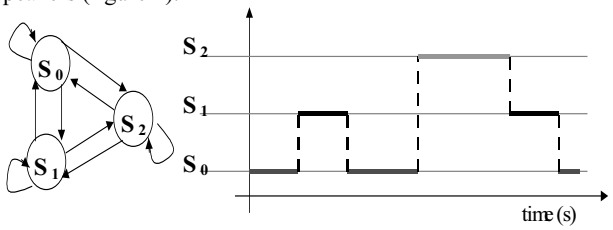

Figure 1: LIA/HMM modeling of the conversation

The speaker segmentation system is applied separately on each of the four acoustic classes detected by the acoustic segmentation described in section 2.1. 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 [5] 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 [7] is used for the re-segmentation phase while a LIA variant optimized for an adaptation on a very short segment [1] 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[2] 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 [6].

2.3 The CLIPS System

The CLIPS system [6] 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 2.1. A BIC [3] 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 (NSp) is estimated using a

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

penalized BIC (Bayesian Information Criterion), in contrast with the last year CLIPS segmentation system which used a fixed number of speakers [6].

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.

3. 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.

3.1 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.

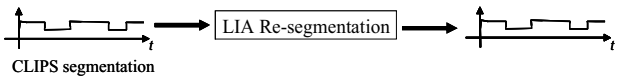

Figure 2: ELISA piped system

3.2 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".

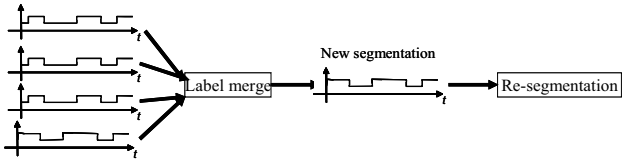

Figure 3 ELISA merge system

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.

4. EXPERIMENTS AND RESULTS

This section presents the results obtained during the NIST RT 03 Evaluation. The audio data provided by NIST for the diarization task consisted of 3 English broadcast news recordings; 30 minutes each containing between 10 and 27 speakers.

	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 1 Experimental results on RT 2003 data

In order to measure the accuracy of the segmentation, the RT scoring system computes an optimum one-to-one mapping between the reference and the hypothesis speakers[3]. An overall time-based speaker diarization error score is computed as the fraction of time that is not attributed correctly to a speaker.

[3] 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) hypothesis speaker to which that reference speaker is mapped.

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 1 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 (but their sums are the same). 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.
- 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 3.2: 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[4].

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[5]. The CLIPS algorithm missed only 7% of the real speakers involved in the files (4 speakers out of 57 total speakers).

5. 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. 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 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 on segment durations and probabilities of speakers involved in dialogue, for both ELISA approaches.

6. 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] 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.

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

[4] S. Meignier, D. Moraru, C. Fredouille, L. Besacier, and J.-F. Bonastre, "Benefit of prior acoustic segmentation for speaker segmentation systems". *Paper submitted at ICASSP'04*, Montreal, Canada.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[5] 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.