The NIST 2004 spring rich transcription evaluation : two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation

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

THE NIST 2004 SPRING RICH TRANSCRIPTION EVALUATION: TWO-AXIS MERGING STRATEGY IN THE CONTEXT OF MULTIPLE DISTANT MICROPHONE BASED MEETING SPEAKER SEGMENTATION

Corinne Fredouille [(2)], Daniel Moraru [(1)], Sylvain Meignier [(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 speaker segmentation approach applied on multiple audio channel meeting recordings in the framework of NIST RT'04s meeting (spring) evaluation campaign. As done for BN data speaker segmentation, the ELISA "meeting" system involves two speaker segmentation systems developed individually by the CLIPS and LIA laboratories. The main originality consists in a "two-axis" merging strategy, proposed to deal with both multiple expert segmentation outputs and multiple microphone segmentation outputs. While expert merging strategy did not really lead to an improvement of the performance, the individual microphone segmentation merging strategy allowed to provide a global segmentation output from several audio channels (microphones) with acceptable performance. The best system obtained 22.6% of diarization error rate during the NIST RT'04s meeting evaluation.

1. INTRODUCTION

The goal of speaker diarization (or segmentation) 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). In speaker diarization the intrinsic difficulty of the task increases according to the data concerned: (two-speaker) telephone conversations, broadcast news, meeting data.

This paper is related to speaker diarization on meeting data in the framework of NIST 2004 spring "meeting" Rich Transcription (RT'04s) evaluation. Meeting data present three main specificities compared to BN data [1]. Firstly, the speech is fully-spontaneous, highly interactive across participants, and presents a large number of disfluencies as well as speaker segment overlaps. Secondly, the meeting room recording conditions associated with distant (table) microphones lead to noisy recordings, including background noises, reverberations and distant speakers. Thirdly, meeting conversations are recorded in smart spaces where multiple sensors are used. Thus, the speaker diarization system has to treat multiple speech channels coming from multiple microphones. The choice of an efficient merging strategy in order to discard the irrelevant information is then an important issue. This last point is the core problem addressed in this paper.

Section 2 of this paper presents the two ELISA speaker diarization systems. *Section 3* describes the strategies used to specifically treat meeting data by merging multiple microphone segmentation outputs and optionally multiple experts. *Section 4* presents the experimental protocols and results. Finally, *section 5* concludes this work.

2. SPEAKER SEGMENTATION SYSTEMS

Two speaker segmentation systems are involved in this work, developed individually by the CLIPS and LIA laboratories in the framework of the ELISA consortium [2]. Both of them participated at the Rich Transcription 2003 evaluation campaign (RT'03) for the speaker segmentation task on broadcast news data [3].

No particular tuning has be done on both systems to participate at RT'04s evaluation campaign except the use of a speech/non speech segmentation as a preliminary phase to deal with the specificities of meeting data.

2.1 Speech/non speech segmentation

The speech/non speech segmentation system consists 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 is set to 0.5s.

2.2. The LIA System

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

The speaker segmentation system is applied on the speech segments detected by the speech/non speech segmentation described in section 2.1.

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. This speaker detection process is then followed by a re-segmentation phase (iterative adaptation and decoding process) which allows to refine speaker segmentation. The entire speaker segmentation process is largely described in [3][4].

Concerning the front end processing, 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. No frame removal or any coefficient normalization is applied.

2.3 The CLIPS System

The CLIPS system is based on a BIC [5] (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 entire speaker segmentation process is largely described in [3][4]. Finally, the re-segmentation phase of the LIA system is also applied on the CLIPS segmentation for refinement[1].
Like the LIA system, the CLIPS system is applied on the speech segments detected by the speech/non speech segmentation.

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 the energy. No frame removal or any coefficient normalization is applied.

3. MEETING SPEAKER SEGMENTATION STRATEGIES

Since meetings are generally recorded with multiple distant microphones, the speaker segmentation task differs greatly from other domains like broadcast news or telephone conversations. Indeed, speaker segmentation system has to deal with multiple speech signals (from the different distant microphones) when the objective is to provide a single meeting speaker segmentation output. Moreover, according to the distant microphone position in the table, the quality of signal may hugely differ from one microphone to another. For instance, the main speaker utterances may be caught by one or two distant microphones while the other microphones mainly provide background voices, long silence, or background noise only.

To deal with these different issues, two cooperative merging strategies are presented in this paper. The first one, called "expert merging strategy" aims at merging segmentations provided by different experts (two experts in this paper). It is applied independently on each recording issued from a distant microphone. The second one, called "Individual Microphone Segmentation Merging strategy (IMSM)", is used to produce a single speaker segmentation output from those obtained on each individual distant microphone. The application of both strategies, also referred as two merging axes – horizontal and vertical –, is illustrated on figure 1.

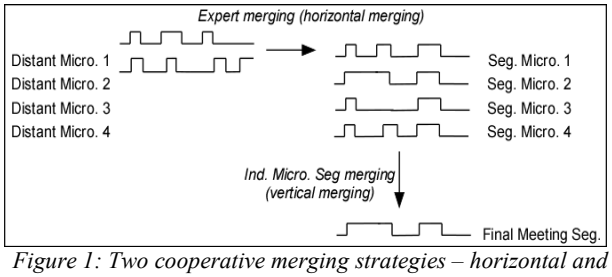

Figure 1: Two cooperative merging strategies – horizontal and vertical merging combination

[1] This combination of CLIPS system and LIA re-segmentation phase was also proposed as a merging strategy during RT'03 evaluation [4] and obtained the best performance over all the participants with 12,88% of speaker diarization error rate.

3.1 Expert Merging Strategy

The idea of this strategy is to merge the segmentations issued from two experts – CLIPS and LIA systems – computed independently on a given distant microphone.

This strategy was already used by the LIA and CLIPS labs for the RT'03 speaker segmentation evaluation campaign on broadcast news data [4]. It relies on a frame based decision which consists in grouping the labels proposed by both the systems at the frame level before applying a re-segmentation process (see figure 2).

An example of the label merging approach is illustrated below:

- Frame *i*: Sys1="S1", Sys2="T4" → label "S1T4",
- Frame *i+1* : Sys1="S2", Sys2="T4" → label "S2T4"

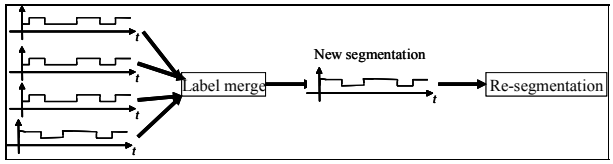

Figure 2: Expert merging strategy

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

- Virtual speakers that have a large amount of data assigned. These speakers could be considered the correct hypothesis speakers;
- Virtual speakers generated by only one of the two 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 zones of indecision.

Based on these considerations, the LIA re-segmentation is then applied on the merged segmentation. During this iterative process, 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 iteration.

After the first iteration the number of speakers is already drastically reduced 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 behaviour 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 label merging phase and since no clustering is applied before re-segmentation, it leads to a "true" speaker split in two virtual speakers.

3.2 Individual Microphone Segmentation Merging Strategy

The goal of this strategy is to merge the multiple distant microphone segmentations in a single meeting speaker segmentation output. Since no single signal is representative of the overall meeting, this strategy must rely on some segment selection rules over the multiple distant microphone speaker segmentations.

In this way, a specific merging algorithm is proposed in this paper. Developed by the LIA and CLIPS labs, it relies on an iterative process which aims at detecting the longest speaker interventions over the set of distant microphone segmentations. This algorithm consists in 3 steps :

- Step 1: selecting the longest speaker intervention over all microphone segmentation outputs taken separately. The longest speaker intervention means all the segments (contiguous or not) attributed to the speaker over a specific microphone segmentation. These segments are definitely attributed to a new speaker in the resulting segmentation.
- Step 2: deleting in each distant microphone segmentation all the segments attributed to the new speaker at the end of step 1.
- Step 3: verifying the presence of not selected segments over all the distant microphone segmentations. If segments are still present and their total length is greater than 30s, then back to step 1 for a new iteration, else stop the process and assign the segments to a last speaker label (this last speaker can be seen as a "trash" speaker related to all the short remaining segments).

One rule is used during this iterative process :

- if the longest speaker intervention selected during step 1 is longer than 60% of the overall signal duration, it is not considered (unless it is the last available intervention). This rule aims at discarding some very long speaker segmentation outputs, which may result from poor individual microphone segmentations (the badness of an individual microphone segmentation may be due, for instance, to the major presence of background voice/noise over the microphone signal, involving a large rate of speech/non speech segmentation errors).

4. EXPERIMENTS AND RESULTS

4.1 Evaluation protocols

RT'04s meeting evaluation campaign [6], proposed two main tasks: speech-to-text transcription (STT) and/or speaker segmentation (so called diarization). For both tasks, different microphone conditions were available: multiple distant microphones, single distant microphone and individual head microphone (the latter was available for STT only).

This paper addresses only speaker segmentation over multiple distant microphones. This section describes the evaluation protocols used to measure the performance, presents some results and discusses the behaviour of the two axis merging strategy.

Scoring

In order to measure performance, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs is computed, followed by a time based speaker segmentation error rate. This scoring, proposed by NIST, is described in details in the RT'04s evaluation plan [7]. Speaker segmentation performance is expressed in terms of speaker diarization error, comprising missed and false alarm speaker errors as well as speaker segmentation errors.

NB: In this paper, the areas of overlap between speaker utterances are not scored.

Database

Since this work was done in the context of RT'04s evaluation campaign, two meeting corpora are available, named in this paper *Dev* corpus for the development of systems and *Eva* corpus for the evaluation. Both of them are composed of two 10mn meeting excerpts recorded over four different sites (CMU, ICSI, LDC, and NIST). Table 1 provides some details on the different corpora, including, for each meeting excerpt, the number of available distant microphones. For each distant microphone, their position in the meeting room is available as further information and may be used to help speaker segmentation process. Nevertheless, approaches presented in this paper do not take advantage of this kind of information.

Finally, as for any speaker segmentation evaluation, no prior information about the number of speakers and their identity is available.

Dev		*Eva*	
Meetings	micro nb	Meetings	micro nb
CMU_20020319-1400	1	*CMU_20030109-1530*	1
CMU_20020320-1500	1	*CMU_20030109-1600*	1
ICSI_20010208-1430	6	*ICSI_20000807-1000*	6
ICSI_20010322-1450	6	*ICSI_20011030-1030*	6
LDC_20011116-1400	7	*LDC_20011121-1700*	10
LDC_20011116-1500	8	*LDC_20011207-1800*	4
NIST_20020214-1148	7	*NIST_20030623-1409*	7
NIST_20020305-1007	6	*NIST_20030925-1517*	7

Table 1: Number of distant microphones for each meeting of Dev and Eva corpora.

4.2 Results

Tables 2 and 3 provide the experimental results obtained on *Dev* and *Eva* corpora for the task of multiple distant microphone speaker segmentation. These results, expressed in terms of speaker diarization error rates, are given for three different systems:

- LIA+IMSM: the LIA speaker segmentation system applied on each individual distant microphones and followed by the Individual Microphone Segmentation Merging (IMSM) process;
- CLIPS+IMSM: the same process is applied using the CLIPS speaker segmentation system followed by the IMSM process;
- Two axis merging: application of the expert merging strategy on the LIA and CLIPS segmentations followed by the IMSM process.

These results show:

- important differences in performance between the LIA and CLIPS systems on a same meeting file (e.g. 14.1% vs 53.4% for *CMU_20020320-1500* on *Dev* corpus and 37.9% vs 19.1% for *ICSI_20000807-1000* on *Eva* corpus);
- important differences in performance between the meetings (e.g. 7.4% vs 54.1% for the LIA between *LDC_20011116-1400* and *NIST_20020305-1007* on *Dev* corpus);
- a significant difference of performance between *Dev* and *Eva* corpora (from 22.6% for the best overall error rate on *Eva* vs 28.3% on *Dev*) as well as a different behaviour of systems between corpora (LIA system is the best one on *Dev* and CLIPS system the best one on *Eva*);

- a small performance improvement observed with the two axis merging strategy compared to the individual systems, and only on few meeting files, (e.g. 25.3% for two axis merging vs 28.4% for the LIA and 26.7% for the CLIPS for *LDC_20011207-1800*). Nevertheless, no gain is reached on the overall performance, compared to the best individual system.

4.3 Discussion

According to the difficulty of the task (compared to broadcast news or conversational telephone data), the performance obtained by the various systems is quite satisfying, especially on *Eva* corpus: 22.6% for the best system, to be compared with 12,88%[1] obtained on BN data during RT'03.

Nevertheless, the "expert merging strategy" applied individually on each individual microphone ("two axis merging") does not provide additional performance gain compared to the best system. This result differs from RT'03 ones [4] where a 16% relative decrease of the diarization error was observed (from 16,90% for the best individual system to 14,24% for the expert merging based system). Moreover, the behaviour of this strategy greatly depends on the quality of individual segmentations, when themselves are dependent on the quality of each stream caught by each individual microphone. One explanation of the disappointing behaviour of the expert merging strategy may be that each expert is applied separately on a missing data file (i.e. on each individual microphone recording). Thus, the performance of the two experts may be very different for a same meeting file, which is a well known drawback in fusion (it is generally well accepted that an efficient fusion must be done between experts that have not too large differences in terms of performance).

Table 4 shows the differences between the microphones taken independently, on two different meeting examples[2]. In the first example (*LDC_20011116-1500)*, the result shows a large variability in terms of speaker error rates between the microphones (d3, d5, d6…). Contrarily, regarding the speech/non speech detection, a small variability between the microphones is noted. On this same meeting, the overall score is very close to the best individual microphone result, which performs quite well. The second example (*NIST_20020305-1007*) shows an inverse behaviour: comparable and quite reasonable speaker error rates over the set of microphones vs. high missed speech error rates with a large variability between the microphones. The differences observed between the meetings show the difficulty to define an efficient merging strategy.

To summarize, some comments could be proposed regarding the results:

- If one microphone is able to catch the information from all the speakers (d2, *LDC_20011116-1500* for example), this microphone could be used alone achieving good performance (14,5% of diarization error on the previous example to be compared with 12,88 % on BN data);
- If the information is present simultaneously on different microphones (with different signal qualities), the fusion process is disturbed, since it is not able to group two (or more) parts of a given speaker detected on different microphones together;
- To take advantage of the multiple microphones, it is necessary to focus on the useful information/speakers present in each recording, i.e. the speech/non speech process should delete the far speakers (low SNR parts, background voices…).

[2] Speaker diarization error rates provided in table 4 for each distant microphone are computed by mapping each individual microphone segmentation to the corresponding single meeting reference segmentation.

[3] The speaker error rate is computed only on well detected speech segments (speech segments present both in the reference and in the system output).

5. CONCLUSION

We have presented the ELISA speaker segmentation approach applied on meeting speech data for NIST RT'04s (spring) evaluation campaign. The best system obtained 28.3% of diarization error on the development corpus (*Dev*) and 22.6% on the evaluation corpus (*Eva*), to be compared with the 12,88% obtained on BN data during NIST RT'03 evaluation. A simple "two-axis merging" strategy was proposed to treat multiple expert segmentation outputs and multiple microphone segmentation outputs. While expert merging strategy did not really lead to an improvement of the performance, the individual microphone segmentation merging strategy allowed to provide a global segmentation output from several audio channels (microphones) with acceptable performance.

To be efficient when the speaker voices are differently caught by the microphones, our simple merging strategy needs microphone independent segmentations focused only on the well caught speakers (the background/far speakers should be suppressed).

Despite the simplicity of the merging strategy proposed in this paper, the ELISA primary system presented to the RT'04s (spring) meeting evaluation obtained the best performance on the speaker diarization task.

6. REFERENCES

[1] http://www.nist.gov/speech/test_beds/mr_proj/

[2] 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.

[3] 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.

[4] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The ELISA consortium approaches in Broadcast News speaker segmentation during the NIST 2003 Rich Transcription evaluation". *ICASSP'04,* Montreal, Canada, May 2004.

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

[6] http://nist.gov/speech/tests/rt/rt2004/spring/

[7] http://nist.gov/speech/tests/rt/rt2004/spring/documents/rt04s-meeting-eval-plan-v1.pdf

	Speaker diarization error (in %)		
***Dev* Meeting Corpus**	**LIA+ IMSM**	**CLIPS+ IMSM**	**Two axis merging**
CMU_20020319-1400	58.5	42.4	47
CMU_20020320-1500	14.1	53.4	52.7
ICSI_20010208-1430	16.9	25.9	18.9
ICSI_20010322-1450	26.5	26.8	27.1
LDC_20011116-1400	7.4	7.5	7.5
LDC_20011116-1500	13.9	16.4	18.1
NIST_20020214-1148	30.8	31.4	33.3
NIST_20020305-1007	54.1	36.8	35.5
Overall (miss. and fa non speech err.=5.6%)	**28.3**	**29.9**	**29.8**

Table 2: Performance (in terms of speaker diarization error rate) of individual speaker segmentation systems (LIA and CLIPS) applied on each distant microphones followed by Individual Microphone Segmentation Merging (IMSM) Strategy and of two axis merging strategy based system. Performance given for each Dev corpus meeting signal and for the overall.

	Speaker diarization error (in %)		
***Eva* Meeting Corpus**	**LIA+ IMSM**	**CLIPS+ IMSM**	**Two axis merging**
CMU_20030109-1530	20.8	39.8	41.2
CMU_20030109-1600	13.7	17.8	18.8
ICSI_20000807-1000	37.9	19.1	17.2
ICSI_20011030-1030	52.1	44.2	42.2
LDC_20011121-1700	16.5	7.7	18.0
LDC_20011207-1800	28.4	26.7	25.3
NIST_20030623-1409	10.3	13.9	10.6
NIST_20030925-1517	22.9	22.7	23.8
Overall (miss. and fa non speech err.=7%)	**24.4**	**22.6**	**23.4**

Table 3: Performance (in terms of speaker diarization error rate) of individual speaker segmentation systems (LIA and CLIPS) applied on each distant microphones followed by Individual Microphone Segmentation Merging (IMSM) Strategy and of two axis merging strategy based system. Performance given for each Eva corpus meeting signal and for the overall.

	Error rates (in %)			
	LDC_20011116-1500		*NIST_20020305-1007*	
Micro	**Mis+fa err. rate**	**Speaker err. Rate**	**Mis+fa err. rate**	**Speaker err. rate**
d1	3.7	18.6	34.4	22.9
d2	4.9	9.6	21.8	26.3
d3	4.9	47.9	XX	XX
d4	7.4	11.6	20	29
d5	4.0	48.5	36.2	13.9
d6	3.1	48.5	29.2	19.3
d7	4.5	48.3	25.2	16.2
d8	7.3	47.6	XX	XX
IMSM	**2.5**	**11.4**	**10.2**	**43.9**

Table 4: two examples of Individual Microphone Segmentation Merging (IMSM) strategy behaviour for the LIA+IMSM system.