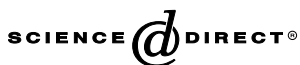




Available online at www.sciencedirect.com



Computer Speech and Language xxx (2005) xxx–xxx

COMPUTER
SPEECH AND
LANGUAGE

www.elsevier.com/locate/csl

Step-by-step and integrated approaches in broadcast news speaker diarization

Sylvain Meignier ^{a,c}, Daniel Moraru ^b, Corinne Fredouille ^{a,*},
Jean-François Bonastre ^a, Laurent Besacier ^b

^a *Laboratoire Informatique d'Avignon (LIA)/CNRS, Department of Computing, University of Avignon, BP1228, 84911 Avignon Cedex 9, France*

^b *CLIPS, IMAG (UJF & CNRS), BP 53, 38041 Grenoble Cedex 9, France*

^c *LIUM/CNRS, Université du Maine, Avenue Laennec, 72085 Le Mans Cedex 9, France*

Received 2 November 2004; received in revised form 1 August 2005; accepted 3 August 2005

Abstract

This paper summarizes the collaboration of the LIA and CLIPS laboratories on speaker diarization of broadcast news during the spring NIST Rich Transcription 2003 evaluation campaign (NIST-RT'03S). The speaker diarization task consists of segmenting a conversation into homogeneous segments which are then grouped into speaker classes.

Two approaches are described and compared for speaker diarization. The first one relies on a classical two-step speaker diarization strategy based on a detection of speaker turns followed by a clustering process, while the second one uses an integrated strategy where both segment boundaries and speaker tying of the segments are extracted simultaneously and challenged during the whole process. These two methods are used to investigate various strategies for the fusion of diarization results.

Furthermore, segmentation into acoustic macro-classes is proposed and evaluated as a priori step to speaker diarization. The objective is to take advantage of the a priori acoustic information in the diarization process. Along with enriching the resulting segmentation with information about speaker gender,

* Corresponding author. Tel.: +33 4 90 84 35 78; fax: +33 4 90 84 35 01.

E-mail addresses: sylvain.meignier@univ-lemans.fr (S. Meignier), daniel.moraru@imag.fr (D. Moraru), corinne.fredouille@lia.univ-avignon.fr (C. Fredouille), jean-francois.bonastre@lia.univ-avignon.fr (J.-F. Bonastre), laurent.besacier@imag.fr (L. Besacier).

2

S. Meignier et al. / Computer Speech and Language xxx (2005) xxx–xxx

25 channel quality or background sound, this approach brings gains in speaker diarization performance
26 thanks to the diversity of acoustic conditions found in broadcast news.

27 The last part of this paper describes some ongoing works carried out by the CLIPS and LIA laboratories
28 and presents some results obtained since 2002 on speaker diarization for various corpora.

29 © 2005 Elsevier Ltd. All rights reserved.

30 *Keywords:* Speaker indexing; Speaker segmentation and clustering; Speaker diarization; E-HMM; Integrated approach;
31 Step-by-step approach

32

33 1. Introduction

34 The design of efficient indexing algorithms to facilitate the retrieval of relevant information is
35 vital to provide easy access to multimedia documents. Until recently, indexing audio-specific
36 documents such as radio broadcast news or the audio channel of video materials mostly con-
37 sisted of running automatic speech recognizers (ASRs) on the audio channel in order to extract
38 syntactic or higher level information. Text-based information retrieval approaches were then ap-
39 plied to the transcription issued from speech recognition. The transcription task alone repre-
40 sented one of the main challenges of speech processing during the past decade (see the
41 DARPA workshop proceedings at [Darpa speech recognition evaluation workshop](#)) and no spe-
42 cific effort was dedicated to other information embedded in the audio channel. Progress made in
43 broadcast news transcription (Kim et al., 2003; Nguyen and Xiang, 2004) shifts the focus to a
44 new task, denoted “Rich Transcription” (NIST-RT03S, 2003), where syntactic information is
45 only one element among various types of information. At the first level, acoustic-based infor-
46 mation like speaker turns, the number of speakers, speaker gender, speaker identity, other
47 sounds (music, laughs) as well as speech bandwidth or characteristics (studio quality or tele-
48 phone speech, clean speech or speech over music) can be extracted and added to syntactic infor-
49 mation. At the second level, information directly linked to the spontaneous nature of speech,
50 like disfluencies (hesitations, repetitions, etc.) or emotion is also relevant for rich transcription.
51 On a higher level, linguistic or pragmatic information such as named entity or topic extraction
52 for instance is particularly interesting for seamless navigation or multimedia information retrie-
53 val. Finally, some types of information extraction relevant to document structure do not fall
54 exactly into one category; for example, the detection of sentence boundaries can be based on
55 acoustic cues but also on linguistic ones.

56 This paper concerns information extraction on the first level described above. It is mainly
57 dedicated to the detection of speaker information, such as speaker turns, speaker gender, and
58 speaker identity. These speaker-related tasks correspond to speaker segmentation and clustering,
59 also denoted speaker diarization in the NIST rich transcription (RT) evaluation campaign
60 terminology.

61 The speaker diarization task consists of segmenting a conversation involving multiple speakers
62 into homogeneous parts which contain the voice of only one speaker, and grouping together all
63 the segments that correspond to the same speaker. The first part of the process is also-called
64 speaker change detection while the second one is known as the clustering process. Generally,
65 no prior information is available regarding the number of speakers involved or their identities.

66 Estimating the number of speakers is one of the main difficulties for the speaker diarization task.
67 To summarize, this task consists of:

- 68 • finding the speaker turns,
- 69 • grouping the speaker-homogeneous segments into clusters,
- 70 • estimating the number of speakers involved in the document.

71
72 Classical approaches for speaker diarization (Siu et al., 1992; Wilcox et al., 1994; Siegler et al.,
73 1997; Gauvain et al., 1998; Chen and Gopalakrishnan, 1998) deal with these three points succes-
74 sively: first finding the speaker turns using by example the symmetric Kullback Leibler (KL2), the
75 generalized likelihood ratio (GLR), or the Bayesian information criterion (BIC) distance ap-
76 proaches, then grouping the segments during a hierarchical clustering phase, and finally estimat-
77 ing the number of speakers a posteriori. If this strategy presents some advantages like dealing with
78 quite long and pure segments for the clustering, it also has some drawbacks. For example, knowl-
79 edge issued from the clustering (like speaker-voice models) could be very useful to estimate seg-
80 ment boundaries as well as to facilitate the detection of other speakers. Contrasting with this
81 *step-by-step* strategy, an *integrated* approach, for which the three steps involved in speaker diar-
82 ization are performed simultaneously, uses all the information currently available for each of the
83 subtasks (Meignier et al., 2001; Ajmera and Wooters, 2003). The main disadvantage of the *inte-*
84 *grated* approach lies in the need to learn robust speaker models using very short segments (rather
85 than a cluster of segments as in classical approaches), even though the speaker models get refined
86 along the process. Mixed strategies are also proposed (Wilcox et al., 1994; Reynolds et al., 2000;
87 Moraru et al., 2004), where classical *step-by-step* segmentation and clustering are first applied and
88 then refined using a “re-segmentation” process during which the segment boundaries, the segment
89 clustering and sometimes the number of speakers are challenged jointly.

90 In addition to the intrinsic speaker diarization subtasks presented above (denoted p1 in the list
91 below), various problems need to be solved in order to segment an audio document into speakers,
92 depending on the environment or the nature of the document:

- 93 • to identify the speaker turns and the speaker clusters, and to estimate the number of speakers
94 involved in the document, without any a priori information (p1);
- 95 • to be able to process speech documents as well as documents containing music, silence, and
96 other sounds (p2);
- 97 • to be able to process spontaneous speech with overlapping voices of speakers, disfluencies, etc.
98 (p3).

99 The NIST’02 speaker recognition evaluation provided an overview of the performance that can be
100 obtained for:

- 101 • conversational telephone speech, involving two speakers and a single acoustic class of signals;
- 102 • broadcast news data which often includes various qualities or types of signal (such as studio/
103 telephone speech, music, speech over music, etc.);
- 104 • meeting room data in which speech is more spontaneous than in the previous cases, and presents
105 several distortions due to distant microphones (e.g., table microphone) and noisy environment.

106
107 Table 1 shows the various classes of problems encountered in each situation (p1, p2, and p3).
108 The increasing difficulty of the tasks is obviously due to their novelty (the last two tasks were
109 introduced for the 2002 evaluation campaign) but also and mainly to the accumulation of prob-
110 lems described in the previous paragraph.

111 Since 2001, two members of the ELISA Consortium, CLIPS and LIA, have been collaborating
112 in order to participate in the yearly evaluation campaigns for the task of speaker segmentation/
113 diarization: NIST'01 (2001) (LIA only), NIST'02 (2002), NIST-RT'03S (2003), and NIST-RT'04S
114 (2004). Since speaker diarization may also be useful for indexing and segmenting videos, CLIPS
115 has also participated in experiments in the last three TREC VIDEO evaluations (Smeaton et al.,
116 2003) since 2002 (Quénot et al., 2002; Quénot et al., 2003).

117 The ELISA Consortium was originally created by ENST, EPFL, IDIAP, IRISA and LIA in
118 1998 with the aim of promoting scientific exchange between members, developing a common
119 state-of-the-art speaker verification system and participating in the yearly NIST speaker recog-
120 nition evaluation campaigns. With the years, the composition of the Consortium has changed
121 and today CLIPS, DDL, ENST, IRISA, LIA, LIUM and the Friburg University are members.
122 Since 1998, the members of the Consortium have participated in the NIST evaluation cam-
123 paigns in speaker verification; a comparative study of the various systems presented in the
124 1999, 2000 and 2001 campaigns can be found in ELISA (2000) and Magrin-Chagnolleau
125 et al. (2001).

126 This paper presents an overview of this long-term collaboration by investigating two main
127 issues. Firstly, the relative advantages of the classical *step-by-step* approach as well as of a
128 more original *integrated* strategy are discussed (this part of the work can be linked to the
129 “p1” point mentioned above: the intrinsic tasks of speaker diarization). Several fusion strate-
130 gies that use the advantages of both approaches are also proposed. The second issue addressed
131 in this paper concerns the nature of the audio documents to be segmented (issue denoted as
132 “p2”). This part of the work is more precisely dedicated to speaker diarization of broadcast
133 news data. The interest of applying an acoustic macro-class segmentation process before
134 speaker segmentation (in order to divide the audio file into bandwidth- or gender-homoge-
135 neous parts) is discussed.

136 This paper is organized as follows: Section 2 is devoted to the description of systems. The
137 acoustic macro-class segmentation process and the two speaker diarization techniques are de-
138 scribed successively. Section 3 focuses on the fusion of the two approaches. Performance of the
139 various systems is presented and discussed in Section 4. All the experimental protocols and data
140 are issued from the NIST-RT'03S development and evaluation corpora (except for some results
141 on meeting data reported in Section 5, issued from the NIST-RT'04S meeting data evaluation
142 (NIST-RT'04S, 2004; Fredouille et al., 2004)). Section 5 presents ongoing work on meeting data

Table 1
Increasing difficulty of the tasks

Task	Telephone	Broadcast news	Meeting
Diarization error rate	5.7	26.4	30.1
Problems involved	p1 (but with fixed number of speakers)	p1 + p2	p1 + p2 + p3

Best results for the speaker diarization task in the NIST'02 speaker recognition evaluation.

143 and integration of a priori knowledge into a speaker diarization system. Finally, concluding re-
 144 marks are made in Section 6.

145 **2. Speaker diarization approaches**

146 Two different speaker diarization systems are proposed in this paper and described in the next
 147 sections. They were developed individually by the CLIPS and LIA laboratories in the framework
 148 of the ELISA consortium (Moraru et al., 2003; Moraru et al., 2004). The CLIPS system relies on a
 149 classical step-by-step strategy. It involves a distance based detector strategy (Delacourt and Wel-
 150 kens, 2000) followed by a hierarchical clustering. This approach will be denoted as *step-by-step*
 151 strategy in the rest of this paper. The second system developed by the LIA follows an integrated
 152 strategy. It is based on a HMM and will be denoted as *integrated* strategy in this paper.

153 As illustrated in Fig. 1, both systems use an acoustic macro-class segmentation as a preliminary
 154 phase. During this acoustic segmentation, the signal is first divided into four acoustic classes
 155 according to different conditions based on gender and wide/narrow band detection. Then, the
 156 (CLIPS and LIA) diarization systems are individually applied on each isolated acoustic class. Fi-
 157 nally, the four resulting segmentation outputs are merged and consolidated through a re-segment
 158 at ion phase. The separate application of the speaker diarization systems on each acoustic class as-
 159 sumes that a particular speaker is associated with one of them only. Nevertheless, the re-segmen-
 160 tation process allows to question the relationship between a speaker and a unique acoustic class.

161 Both diarization approaches and acoustic segmentation were developed independently before
 162 investigating different strategies for combining the systems. Therefore, the settings of each of
 163 them, like acoustic features or learning methods, may differ but come from experiments conducted
 164 over a common development corpus (see Section 4.1).

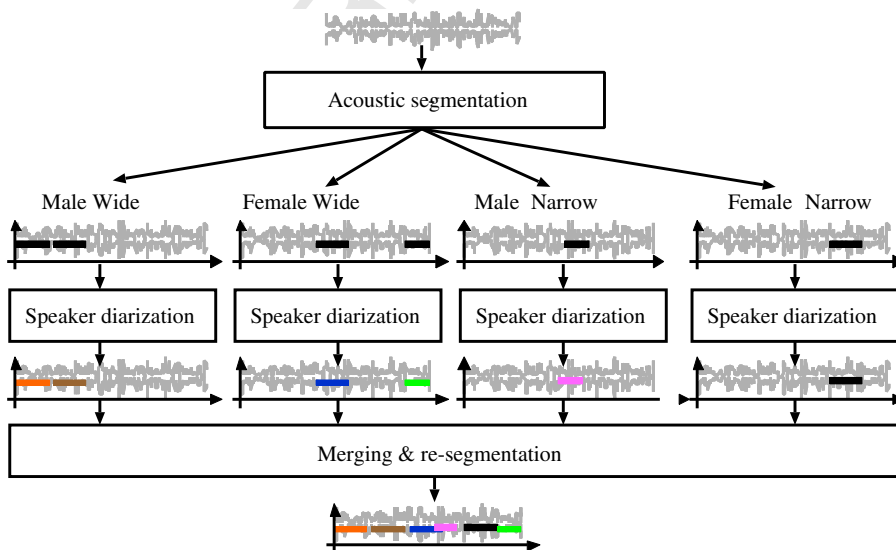


Fig. 1. Overview of the speaker diarization strategy.

165 2.1. Acoustic macro-class segmentation

166 Segmenting an audio signal into acoustic classes was mainly introduced to assist ASR systems
 167 within the special context of broadcast news transcription (Hain and Woodland, 1998; Woodland,
 168 2002; Gauvain et al., 2002). Indeed, one of the first objectives of acoustic segmentation was to
 169 provide ASR systems with an acoustic event classification to discard non-speech signal (silence,
 170 music, commercials) and to adapt ASR acoustic models to some particular acoustic environments,
 171 like speech over music, telephone speech or speaker gender. Many papers were dedicated to this
 172 particular issue and to the evaluation of acoustic segmentation in the context of the ASR task.
 173 However, acoustic segmentation may be useful for other tasks linked to broadcast news corpora,
 174 although this is rarely discussed in the literature. In this sense, one of the aims of this work is to
 175 investigate the impact of acoustic segmentation when it is applied as prior segmentation for speak-
 176 er diarization.

177 Speech/non-speech detection is useful for the speaker diarization task in order to avoid music
 178 and silence portions being automatically labeled as new speakers. This is particularly true in the
 179 context of the NIST-RT evaluation in which both miss and false alarm speech errors are taken
 180 into account for the speaker diarization scoring.

181 Moreover, an acoustic segmentation system can be designed to provide a finer classification.
 182 For example, gender and frequency band detection may introduce a priori knowledge in the diar-
 183 ization process. In this paper, the prior acoustic segmentation is done at three different levels:

- 184 • Speech/non-speech.
- 185 • Clean speech/speech over music/telephone speech (narrow band).
- 186 • Male/female speech.

187

188 2.1.1. Hierarchical approach

189 The system relies on a hierarchical segmentation performed in three successive steps as illus-
 190 trated in Fig. 2:

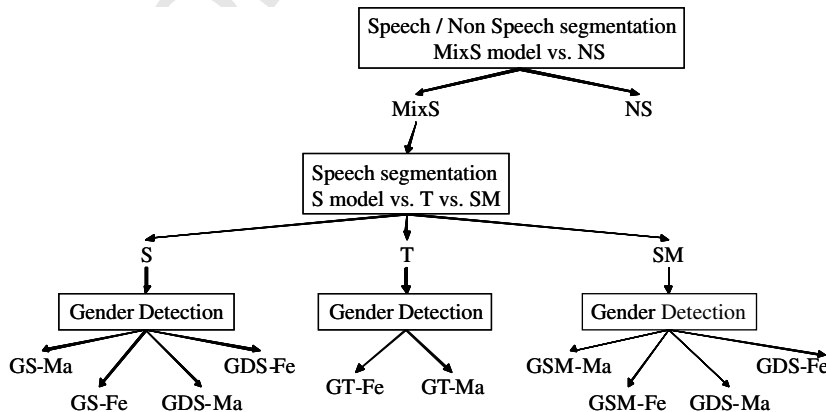


Fig. 2. Hierarchical acoustic segmentation.

- 191 • During the first step, a speech/non-speech segmentation is performed using two models. The
192 first model, *MixS*, represents all the speech conditions while the second one, *NS*, represents
193 the non-speech conditions. Basically, the segmentation process relies on a frame-by-frame
194 best model search. A set of morphological rules is then applied to aggregate frames and to
195 label segments. These rules mainly aim at constraining the duration of segments, by fixing
196 for instance minimum lengths for both speech and non-speech segments. This strategy was
197 preferred to a Viterbi decoding, which tends, in this context, to misclassify non-speech
198 segments.
- 199 • During the second step, a segmentation based on three classes, clean speech (*S* model),
200 speech over music (*SM* model) and telephone speech (*T* model), is performed only on the
201 speech segments detected during the previous segmentation step. All the models involved
202 during this step are gender-independent. The segmentation process is a Viterbi decoding
203 applied on an ergodic HMM, composed of three states (*S*, *T*, and *SM* models). The transi-
204 tion probabilities of this ergodic HMM are learnt on the 1996 HUB 4 broadcast news
205 corpus.
- 206 • The last step is gender detection. According to the label assigned during the previous step, each
207 segment will be identified as female or male speech by the use of models dependent on both
208 gender and acoustic classes. *GT-Fe* and *GT-Ma* models represent female and male telephone
209 speech respectively, *GS-Fe* and *GS-Ma* represent female and male clean speech, while *GSM-*
210 *Fe* and *GSM-Ma* represent female and male speech over music. Two additional models,
211 *GDS-Fe* and *GDS-Ma*, representing female and male speech recorded under degraded condi-
212 tions are also used to refine the final segmentation. The segmentation process described in
213 the previous step is applied here again.
- 214

215 2.1.2. System specifications

216 The signal is characterized by 39 acoustic features computed every 10 ms on 25 ms Hamming-
217 windowed frames: 12 Mel frequency cepstral coefficients (MFCC) augmented by the normalized
218 log-energy, followed by the delta and delta-delta coefficients. The choice of parameters was
219 mainly guided by the literature (Hain and Woodland, 1998).

220 All the models mentioned in the previous section are diagonal Gaussian mixture models
221 (GMMs), trained on the 1996 HUB 4 broadcast news corpus. The *NS* and *MixS* models are char-
222 acterized by 1 and 512 Gaussian components respectively, while the other models are character-
223 ized by 1024 Gaussian components. All these parameters have been chosen empirically following
224 a set of experiments not reported here.

225 2.2. Step-by-step speaker diarization

226 The CLIPS system is a state-of-the-art system based on the speaker change detection followed
227 by a hierarchical clustering. The number of speakers involved in the conversation is automatically
228 estimated. The system uses the acoustic macro-class segmentation described in Section 2.1. The
229 CLIPS diarization is applied individually on every acoustic class as explained in Section 2 and
230 the results are merged at the end. The next subsections will provide a detailed description of every
231 module of the system.

232 2.2.1. Step one: speaker change detection

233 The goal of the speaker change detection is to cut the audio recording into segments contain-
234 ing only the speech of one single speaker. The purpose of the speaker change detection is to
235 find some audio signal discontinuities that will help us distinguish between two consecutive
236 speakers. Those segments will be used as input data for the clustering module. A distance based
237 approach (Delacourt and Welkens, 2000; Chen and Gopalakrishnan, 1998) is used, implying
238 here the GLR. Given two acoustic sequences X and Y we test whether they were produced
239 by the same Gaussian model (the same speaker) M_{XY} or by two different models (two different
240 speakers) M_X and M_Y . This question can be answered using the following GLR ratio,
241 where

$$R_{\text{GLR}}(X; Y) = \log L(X|M_X) + \log L(Y|M_Y) - \log L(XY|M_{XY}). \quad (1)$$

245 A high value of R_{GLR} means that the “two-model hypothesis” is more likely than the “one-model
246 hypothesis”. The first two terms of R_{GLR} is the log-likelihood of the “two-model hypothesis” and
247 the last term is the log-likelihood of the “one-model hypothesis”. A GLR curve is extracted from
248 1.75-s adjacent windows that move along the audio signal. The window size must be small enough
249 to contain only one speaker and large enough to obtain a reliable model. The two windows ad-
250 vance frame by frame. Mono-Gaussian models with diagonal covariance matrices are used to
251 build the GLR curve. The maximum peaks of the curve are the most likely speaker change points.
252 A threshold is then applied on the GLR curve to find speaker changes. The threshold is tuned so
253 that over-segmentation (more speaker changes detected) is provided, as we prefer to detect more
254 segments (which can be further merged by the clustering process) rather than miss speaker
255 changes (which will never be recovered later). The threshold is computed using the mean value
256 of the current curve. Thus, it adapts itself from one file to another.

257 Another system was presented at the NIST'02 speaker recognition evaluation with a priori seg-
258 mentation using fixed length segments (0.75 s). It gave approximately the same performance while
259 being 3 times slower due to the uniform segmentation that leads to far more segments as input of
260 the clustering module.

261 2.2.2. Step two: clustering

262 Now that we have detected the speaker changes, the segments obtained must be grouped (clus-
263 tered) by speaker. The CLIPS clustering uses a hierarchical bottom-up algorithm. A clustering
264 algorithm generally relies on two important elements: the distance between classes and the stop
265 criterion. The distance used is the GLR distance and the stop criterion is the estimated number
266 of speakers. The GLR distance is the GLR ratio (see Eq. (1)) computed between classes rather
267 than consecutive windows. Another difference is that the models used are no longer mono-Gauss-
268 ian as in the speaker change detection but GMMs.

269 First, a diagonal 32 GMM background model is trained on the entire file using a classical EM
270 algorithm. We need a background model to compensate for the lack of data for each speaker. The
271 advantage of using a background model trained on the current file is that it is always suited for the
272 current task. A more complex background model (e.g., 512 GMM diagonal) trained on external
273 data could perform better but makes the speaker diarization system data dependent (the system
274 would work only on the type of data used to train the background model). The size of the model is
275 a good compromise between complexity and performance: beyond 32 Gaussian components

276 we only gain about 0.5% absolute diarization error rate (DER) but we increase the execution
277 time.

278 Segment models are then trained using a linear MAP adaptation (see Meignier et al., 2001, for
279 more details on the linear MAP adaptation) of the background model (means only). GLR dis-
280 tances are then computed between models and the closest segments are merged at each step of
281 the algorithm until N segments are left (corresponding to N speakers detected in the
282 conversation).

283 The number of speakers N is estimated as described in the next section. The clustering is done
284 individually on each acoustic macro-class (namely male/wide, female/wide, male/narrow and fe-
285 male/narrow) and the results are merged in the end.

286 2.2.3. Step three: estimating the number of speakers

287 The algorithm that estimates the number of speakers is based on the penalized BIC (Schwarz,
288 1978).

289 At first, the number of speakers is limited to between 1 and 25. The upper limit usually depends
290 on the recording size.

291 We select the number of speakers (N_{sp}) that maximizes

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

294 where M is the model composed of the N_{sp} speaker models, N_X is the total number of speech
295 frames involved, m is a parameter that depends on the complexity of the speaker models and λ
296 is a tuning parameter empirically set at 0.6. In our case (32 diagonal GMM), m is equal to 64
297 (2 times 32) times the number of acoustic features. The first term is the overall log-likelihood
298 of the data. The second term is used to penalize the complexity of the model. We need the second
299 term because the log-likelihood of the data increases with the number of models (speakers) in-
300 volved in the calculation of $L(X|M)$.

301 Let X_i and M_i be the data and the model of speaker i respectively. The model is obtained by
302 MAP adaptation of the background model over the speaker data as in the previous section. If
303 we make the hypothesis that data X_i depends only on the speaker model M_i then we can prove
304 that the overall log-likelihood of the data becomes

$$L(X|M) = \prod_{i=1}^{N_{\text{sp}}} L(X_i|M_i). \quad (3)$$

307 Results concerning the estimation of the number of speakers will be presented in Section 4.

308 2.2.4. System specifications

309 The signal is characterized by 16 MFCC computed every 10 ms on 20 ms windows using
310 56 filter banks. Then we add the energy parameter. The choice of the number of filters is due
311 to the fact that we work on wide-band data (broadcast news). No frame removal nor
312 coefficient normalization is applied. The parameterization is the same for all system modules
313 of this *step-by-step* diarization system, but is different from that of the *integrated* speaker
314 diarization system and the acoustic segmentation, which were all developed separately in
315 different places.

316 2.3. Integrated speaker diarization

317 The LIA system is based on an evolutive Hidden Markov modeling (E-HMM) of the conversa-
318 tion (Meignier et al., 2000; Meignier et al., 2001; Moraru et al., 2003; Moraru et al., 2004). The
319 HMM is ergodic; all speaker changes are potentially available. Each state of the HMM character-
320 izes a speaker and the transitions model the changes between speakers (Fig. 3). In this iterative ap-
321 proach, both the segmentation and the speaker models are used at each step and are re-evaluated at
322 the next step. During the diarization process, the speakers are detected and added one by one at
323 each iteration. This is the reason why we have named this diarization method *integrated* approach.

324 The speaker diarization system relies on the acoustic macro-class segmentation described in
325 Section 2.1. It is applied separately on each of the acoustic classes detected (e.g., male/wide, fe-
326 male/wide, male/narrow and female/narrow). Finally, the separate speaker diarization outputs
327 are merged followed by a re-segmentation process, described in Section 2.4.

328 2.3.1. Speaker diarization process

329 During the diarization, the HMM is generated using an iterative process which detects and adds
330 a new state (i.e., a new speaker) at each iteration. The speaker detection process is performed in
331 four steps (Fig. 4). An example for a two speaker show is given in Fig. 5.

332 Step 1 – *Initialization*. A first speaker model S_0 is trained on the whole show (broadcast news
show for instance). The segmentation is modeled by a one-state HMM and the whole signal
is assigned to speaker S_0 . At the beginning of the iterative process, S_0 represents all
the speakers of the show. At the end of the process, once all the speakers have been
detected (the $n - 1$ first speakers) and their segments associated with, S_0 should represent
a unique speaker, the last one (the n th speaker).

338 Step 2 – *Adding a new speaker*. A new speaker is extracted from the segments currently labeled
 S_0 representing the speakers that are not detected yet. The new speaker model is trained
using the 3-s region of S_0 that maximize the likelihood ratio between model S_0 and a uni-
versal background model (UBM; Reynolds et al., 2000, see Section 2.3.2). The length of
the initial region must be sufficient to initialize a robust speaker model while containing
one speaker only. This strategy selects the closest data to speaker model S_0 . The 3-s
length is chosen empirically. A corresponding state, labeled S_x (x is the number of itera-
tions), is added to the previous HMM. The transition probabilities are updated according
to a set of rules (more details are given in Section 2.3.2). Finally, the selected 3 s of test are
moved from label S_0 to label S_x in the segmentation hypothesis. Various selection strat-

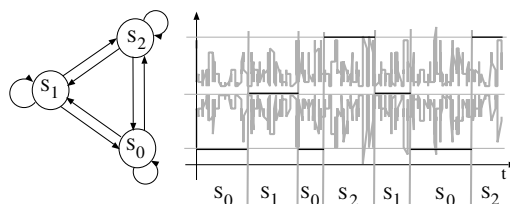


Fig. 3. Integrated approach: evolutive HMM modeling of the conversation and segmentation. Example given for three speakers (S_0, S_1, S_2).

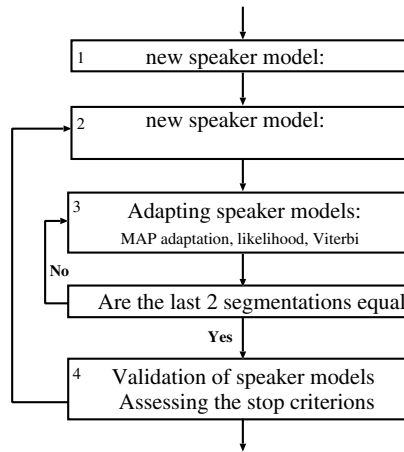


Fig. 4. Integrated approach: different steps of the process.

egies have been tested, involving either the speaker or UBM models. The selection method described here produces the best accuracy in terms of purity of the segments and of speaker diarization error.

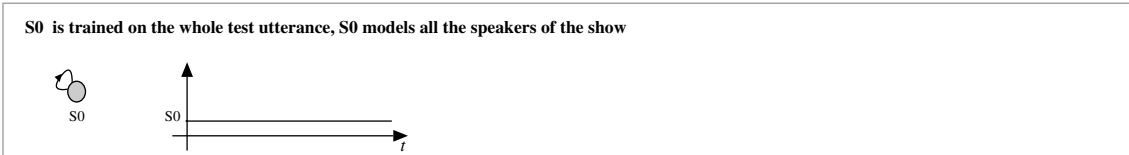
351 Step 3 – *Adapting speaker models*. This phase allows the detection of the segments belonging to a new speaker S_x and the reallocation of the data between all the speakers. First, all the speaker models are adapted according to the current segmentation. Then, Viterbi decoding produces a new segmentation. The adaptation and decoding tasks are performed while the segmentation differs between two successive adaptation/decoding phases. Two segmentations are different when at least a frame is assigned to two different speakers.

357 Step 4 – *Speaker model validation and assessment of the stopping criterion*. The likelihood of the previous solution and the likelihood of the current solution are computed using the current HMM model (for example, the solution with two speakers detected and the current solution with three speakers detected). In order to compare the likelihoods of both solutions, the previous one is rescored using the associated HMM where a non-emitting state is added (i.e., the transition probabilities are set to the same values for both HMM). The stopping criterion is reached when no gain in terms of likelihood is observed or when no more speech is left to initialize a new speaker.

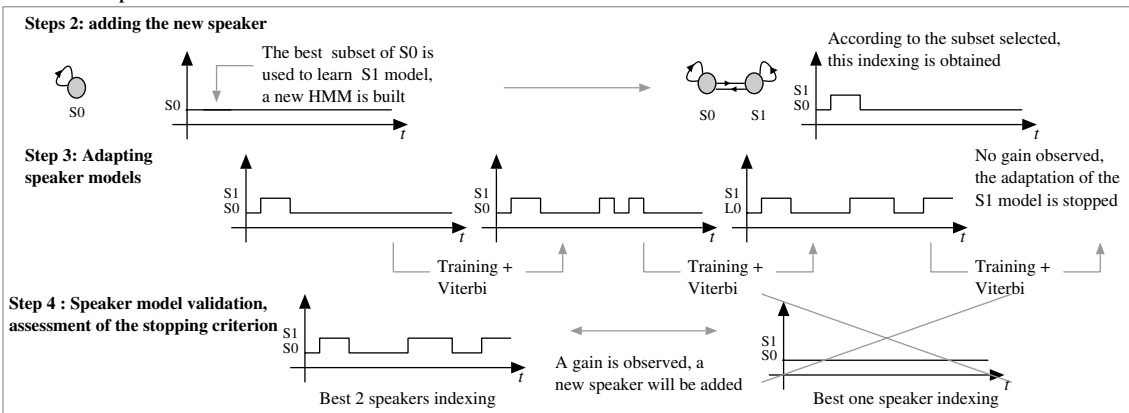
365 During the development, experiments show that two heuristics help to minimize the speaker
 366 diarization error
 367

- 368 • The first one removes the current speaker if the total time of the segments allocated to that
 369 speaker is less than 4 s. Moreover, the 3-s region used for its initialization is never re-employed
 370 in the step 2 and the process continues with the segmentation of the previous iteration.
- 371 • The second one discards the previous speakers from the segmentation if the length of their seg-
 372 ments is lower than the current one. This rule, which forces the detection of the longest speaker
 373 first, is closely related to the evaluation metric used in NIST campaigns where it is more impor-
 374 tant to find the longest speaker segments than the shortest ones.

Step 1: initialization



Iteration 1 : speaker S1



Iteration 2: speaker S2

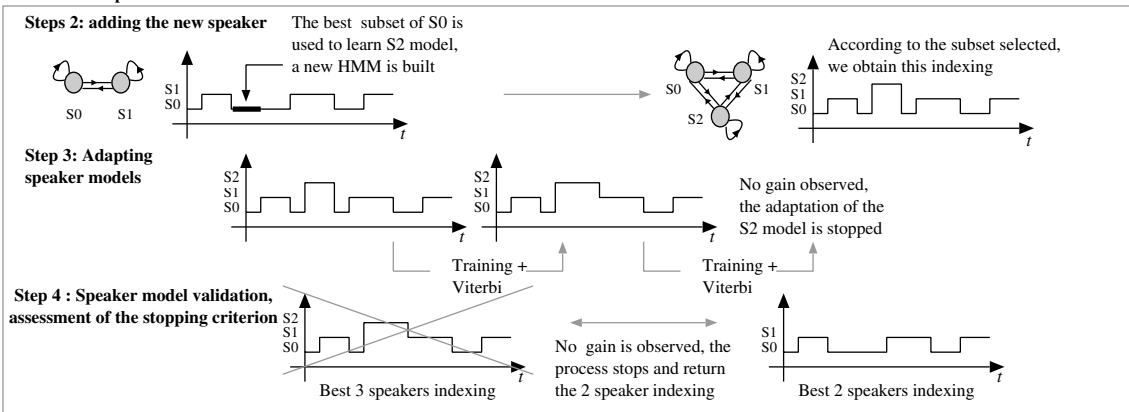


Fig. 5. Integrated approach: diarization example for a two speaker show.

375

376 2.3.2. System specifications

377 The system specifications are set empirically on a development corpus (see Section 4.1). The
378 next paragraphs give some details on the parameterization of the signal, the speaker model adap-
379 tation and the HMM.

380 2.3.2.1. Parameterization. The signal is characterized by 20th order MFCC computed at a 10 ms
381 frame rate using a 20 ms window and the normalized energy. No coefficient normalization is

382 applied; indeed the cepstral mean subtraction (CMS) or the sliding CMS decreases the diarization
383 accuracy.

384 *2.3.2.2. Speaker models.* Speaker models and adaptation techniques used in the E-HMM are sim-
385 ilar to those generally used for automatic speaker recognition. Speaker models are based on
386 GMMs derived from a UBM. Means only are adapted by a MAP technique. The GMMs are
387 composed of 128 Gaussian components with diagonal covariance matrices.

388 The UBM is trained with a classical EM algorithm based on the ML principle on a subset of
389 1996 HUB 4 broadcast news corpus. The UBM learning set is composed of both male and female
390 data and both wide- and narrow-band data. Variance flooring is applied during the training so
391 that variance for each Gaussian is no less than $0.5\times$ the variance of the corresponding UBM
392 Gaussian. A sliding CMS is applied on each training data set before learning; the sliding window
393 is 3 s long. The CMS is performed in order to remove the influence of the various channels (due to
394 the high number of speakers and records in the UBM corpus). Moreover, preliminary experiments
395 have shown an improvement of the speaker diarization accuracy when the UBM features were
396 normalized.

397 The adaptation scheme is based on a variant of MAP developed by the LIA (Meignier
398 et al., 2001). The relative weights of the UBM and the estimate data result from a combi-
399 nation of the UBM and estimated speaker Gaussian weights (respectively $w_i^{\text{UBM}}, w_i^{\text{E}}$ for the
400 Gaussian i) and a priori weights (respectively $\alpha, 1 - \alpha$). The mean i of the speaker model is
401 obtained by

$$\mu_i = \frac{\alpha w_i^{\text{UBM}}}{\alpha w_i^{\text{UBM}} + (1 - \alpha) w_i^{\text{E}}} \mu_i^{\text{UBM}} + \frac{(1 - \alpha) w_i^{\text{E}}}{\alpha w_i^{\text{UBM}} + (1 - \alpha) w_i^{\text{E}}} \mu_i^{\text{E}}. \quad (4)$$

404 Experimentally, α is fixed to 0.2 for the UBM. This setting corresponds to the value that mini-
405 mizes the speaker diarization error over the development corpus.

406 *2.3.2.3. HMM.* The HMM emission probabilities are estimated by computing the mean of the
407 frame-based log likelihoods over a 0.3 s sliding window for each state. This 0.3-s score rate
408 (the systems are generally based on a frame score rate) permits to smooth out local speaker
409 changes and to modify the intrinsic exponential duration law of the states.

410 The HMM transition probabilities are fixed according to the following rules:

- 411 • Each transition probability, $a_{i,i}$ (from state S_i to state S_i) is equal to an *a-priori* value η .
- 412 • Each transition probability, $a_{i,j}$ (from state S_i to state S_j) is equal to

$$a_{i,j} = \frac{(1 - \eta)}{(n - 1)} \quad (5)$$

415 with $i \neq j$ and n is the number of states (i.e., speakers).

416 In this paper, the η value is set to 0.6. This setting corresponds to the value that minimizes the
417 speaker diarization error over the development corpus.
418

419 2.4. *Speaker re-segmentation*

420 The use of a re-segmentation phase at the end of a clustering process was earlier proposed, for
421 example in Wilcox et al. (1994), Gauvain et al. (2001), Reynolds et al. (2000) and Adami et al.
422 (2002). The two main methods are based on GMM/HMM models and make decisions at the
423 frame level:

- 424 • thanks to Viterbi decoding (Wilcox et al., 1994; Gauvain et al., 2001);
- 425 • or over scores computed in a sliding window (Reynolds et al., 2000; Adami et al., 2002).

426 The process can be run iteratively but (Reynolds et al., 2000) has shown that it degrades the
427 performance.

428 The ELISA re-segmentation stage is also based on a Viterbi decoding (similar to the “Adap-
429 ting speaker models” step, described in Section 2.3.1). Firstly, the four gender- and channel-
430 dependent segmentations are merged by simply pooling the segmentations (there is no overlap
431 between sub-segmentations). Secondly, the speaker-model adaptation and Viterbi decoding are
432 performed iteratively. At the end of each iteration, the speakers with less than 4 s of signal are
433 removed.

434 During the re-segmentation process, the parameters are similar to those used for the E-HMM
435 clustering process, except for the model training method. In this case, the classical mean-only
436 MAP adaptation is performed to obtain speaker models (Gauvain and Lee, 1994; Reynolds
437 et al., 2000) instead of the variant MAP technique proposed by the LIA and described in Section
438 2.3.2. The adaptation rate of the means is controlled by the relevant factor (Reynolds et al., 2000)
439 which is experimentally set at 16. Moreover, a tiny gain is obtained over the development corpus
440 when the HMM emission probability score rate is reduced from 0.3 to 0.2 s since this reduction
441 helps to refine the boundaries of the output segmentation.

442 3. **Fusion of systems**

443 Since the NIST 2002 evaluation, CLIPS and LIA have investigated different strategies for com-
444 bining the systems. In this paper, only strategies for broadcast news data are described.¹ Basically,
445 the aim of these strategies is to benefit from the advantages of both speaker diarization ap-
446 proaches, described in previous sections. Two kinds of strategy are proposed: firstly, a hybridiza-
447 tion strategy and secondly, merging various segmentation outputs. The latter is a new way of
448 combining results coming from multiple and unlimited diarization systems.

449 3.1. *Hybridization strategy (“piped” system)*

450 The purpose of this hybridization strategy is to use the results of one system to initialize a sec-
451 ond one. In this paper, the speakers detected by the *step-by-step* system (number of speakers and

¹ The reader is invited to look at Moraru et al. (2003) for telephone strategy.

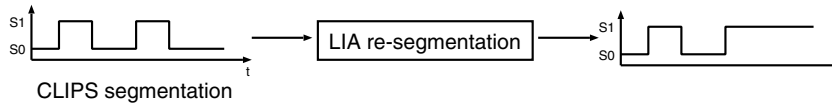


Fig. 6. Example of a piped system.

452 associated audio segments) are inserted in the re-segmentation module of the *integrated* system
 453 (the models are trained using the information provided by the clustering phase) as illustrated in
 454 Fig. 6. This solution associates the advantages of longer and (quite) pure segments, provided
 455 by the *step-by-step* approach, with the HMM modeling and decoding power of the *integrated*
 456 strategy.

457 3.2. Merging strategy (“fusion” system)

458 The aim of the “fusion” system consists of using the segmentation outputs issued from as many
 459 experts as possible. For example, in this paper the total number of experts is four (see Fig. 7): the
 460 *step-by-step* system, the *integrated* system, a variant of the integrated system, and the “piped” sys-
 461 tem (see Section 3.1). The merging strategy relies on a frame-based decision which consists of
 462 grouping the labels proposed by each of the systems at the frame level. An example (for four sys-
 463 tems denoted A, B, C and D) is presented below:

- 464 • Frame i : System A gives the speaker label A_1 ; System B gives B_4 , System C gives C_1 and System
 465 D gives D_1 . $A_1B_4C_1D_1$ is then the merged label.
- 466 • Frame $i + 1$: System A gives A_2 , System B gives B_4 , System C gives C_1 and System D gives D_1 .
 467 $A_2B_4C_1D_1$ is then the merged label.

468 This label merging method generates (before re-segmentation) a large set of potential speakers.
 469 The re-segmentation module of the *integrated* system can be applied on the merged diarization.
 470 Between each adaptation/decoding phase, the potential speakers for whom total time is shorter

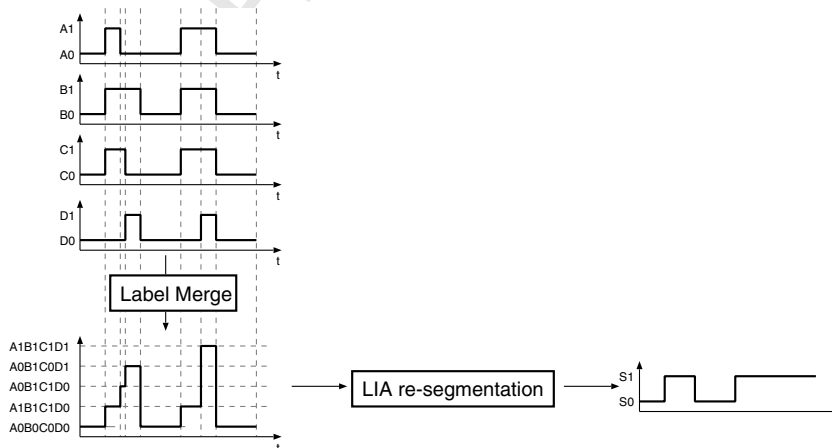


Fig. 7. Example of a merging system.

471 than 3 s are deleted. Indeed, 3 s of signal correspond to the minimal length needed to learn a
472 speaker model. The data of these deleted speakers will further be dispatched between the remain-
473 ing speakers during the next adaptation/decoding phase.

474 4. Experiments and results

475 The experiments were carried out in the framework of the NIST-RT'03S speaker diarization
476 evaluation on American broadcast news (NIST-RT'03S, 2003).

477 4.1. Development and evaluation corpora

478 Following the NIST-RT'03S evaluation campaign, two corpora are available for the speaker
479 diarization task. One of them is used for the development of the systems, which are validated on
480 the second one during a blind evaluation. The development corpus used in this paper is not the
481 official one. It is issued from the RT'02 broadcast news evaluation and was used in the Decem-
482 ber 2002 RT dry-run. Initially, extracted from the HUB-4 evaluation campaign corpus, it is
483 composed of six broadcast news shows of about 10 min each, recorded in 1998 from channels
484 MNB, CNN, NBC, PRI, VOA and ABC. On the other side, the RT'03S evaluation corpus is
485 composed of three 30-min shows recorded in 2001 from channels PRI, VOA and MNB. Some
486 information (averaged size, averaged number of speakers, etc.) related to these corpora are
487 given in Table 2.

488 In this paper, these development and evaluation corpora are named respectively *RT'03S-Dev*
489 and *RT'03S-Eva*. Two additional corpora are used during the experiments. Both of them are de-
490 rived from *RT'03S-Dev* and *RT'03S-Eva* by discarding all the advertisement portions manually
491 before being processed.² Besides, some speech material, not used in the *RT'03S-Dev* and
492 *RT'03S-Eva* corpora, was maintained in the second ones. They are named *ELISA-Dev* (derived
493 from *RT'03S-Dev*) and *ELISA-Eva* (derived from *RT'03S-Eva*) and serve the same role as the ori-
494 ginal corpora, i.e., system development and evaluation purposes. The use of these additional cor-
495 pora during experiments may explain that some results presented in this paper do not correspond
496 exactly to the official *NIST-RT'03S* results.

497 In order to evaluate the accuracy of the acoustic macro-class segmentation, a reference segmen-
498 tation including the different targeted acoustic class (speech/non-speech, gender labels, and tele-
499 phone/non-telephone speech) was necessary. Since NIST does not provide any official reference
500 for the bandwidth classification, the authors have marked their own. Both the boundaries and la-
501 bels were manually identified. This reference segmentation will be referred to as *HandS/NS – Gen-*
502 *der – T/NT* later in this paper.

503 Moreover, it is worth noting that due to the small size of the different corpora, all the results
504 presented in this paper have to be considered with caution.

² Commercials, present in the audio documents, are not scored for the RT'03S evaluation campaign. Nevertheless, their presence during the segmentation process may disturb the systems since they involve additional speakers, entirely irrelevant in the output segmentation.

Table 2

Description of the different corpora, in terms of number of shows, duration average, and speaker number average

Corpus	Number of shows	Duration average (s)	Speaker number average
<i>RT'03S-Dev</i>	6	568	>12 ^a
<i>RT'03S-Eva</i>	3	1534	>19
<i>Elisa-Dev</i>	6	574	12
<i>Elisa-Eva</i>	3	1773	19

^a As commercials are not manually transcribed, the exact number of speakers is unknown.

505 4.2. Evaluation metric

506 The speaker diarization performance is evaluated by comparing the hypothesis segmentation,
507 given by the system, with the reference segmentation provided by NIST. This reference segmen-
508 tation was generated by hand according to a set of rules described in NIST-RT'03S (2003) and
509 NIST (2003).

510 The evaluation metric is based on the NIST speaker diarization metric defined in the NIST-
511 RT'03S evaluation plan (NIST-RT'03S, 2003). It is called the diarization metric, and expressed
512 in terms of diarization error rate (*DER*). It takes three kinds of error into account (named *SE*,
513 *MisE*, *FaE*, respectively in the next sections)

- 514 • A speaker error defined below (*SE*).
- 515 • A missed speaker error relative to a misclassification of speech segments as non-speech seg-
516 ments (*MisE*).
- 517 • A false alarm speaker error relative to a misclassification of non-speech segments as speech seg-
518 ments (*FaE*).

519 To compute the speaker error, the scoring algorithm optimally maps the reference speakers to the
520 hypothesis speakers. Each reference speaker is mapped onto one hypothesis speaker at most and
521 conversely each hypothesis speaker is mapped onto one reference speaker at most. The mapping
522 maximizes the overlap in duration between all pairs of reference and hypothesis speakers. The
523 speaker error is finally expressed as the duration of non-matching zones between reference and
524 hypothesis segments.

525 Concerning the gender- and bandwidth-misclassification errors, they are measured at a frame
526 level by comparing the hypothesis classification with the reference segmentation proposed by
527 the authors *HandS/NS – Gender – T/NT*.

528 4.3. Acoustic macro-class segmentation experiments

529 This section presents the evaluation protocol used to measure the impact of the acoustic macro-
530 class segmentation when combined with speaker diarization and discusses the experimental results
531 obtained in this framework. Different levels of acoustic segmentation granularity are evaluated on
532 both speaker diarization systems:

- 533 • Speech/non-speech classification only (*S/NS*). This segmentation corresponds to the first level
534 of the acoustic macro-class segmentation described in Section 2.1.
- 535 • Segmentation based on speech/non-speech and gender detection (*S/NS-Gender*). This segmen-
536 tation is obtained by merging all the labels *GS-XX*, *GSM-XX*, *GDS-XX* and *GT-XX* yielded by
537 the acoustic macro-class segmentation (see Fig. 2) in a single *XX* label where *XX* represents
538 either *Ma* or *Fe*.
- 539 • Segmentation based on speech/non-speech, gender and telephone/non- telephone speech
540 detection (*S/NS-Gender-T/NT*). *NT* segmentation is obtained by merging all the *GS-XX*,
541 *GSM-XX*, and *GDS-XX* (see Fig. 2) in a single *NT-XX* label where *XX* represents either
542 *Ma* or *Fe*.
- 543 • Segmentation based on speech/non-speech, gender and telephone/clean speech/speech over
544 music/degraded speech (*S/NS-Gender-T/S/MS/DS*). In this segmentation, all the labels yielded
545 by the third level of the acoustic macro-class segmentation system are used (see Fig. 2).

546 For comparison purposes, speaker diarization results based on the reference acoustic macro-class
547 segmentation, *Hand S/NS-Gender-T/NT*, are also presented.

548 4.3.1. Intrinsic performance of acoustic macro-class segmentation

549 Table 3 provides the performance of the acoustic macro-class segmentation on both *RT03S-*
550 *Dev* and *RT03S-Eva* corpora. Some details about the amount of data for each targeted class
551 are reported in Table 4.

552 The speech/non-speech segmentation error is around 4.9% (in terms of duration) compared to
553 4.4% for the best system during the NIST-RT03S evaluation campaign (NIST). The gender detec-

Table 3

Classification error rates made by the acoustic macro-class segmentation system on the *RT03S-Dev* and *RT03S-Eva* sets according to the different classes available on the audio material (speech, non-speech, gender and telephone/non-telephone)

Corpus	Classification error rate (%)			
	Speech	Non-speech	Gender	Telephone/non-telephone
<i>RT03S-Dev</i>	2.3	2.2	1.5	0.09
<i>RT03S-Eva</i>	1.1	3.8	5.5	3.0

Automatic and manual acoustic segmentations are compared at the frame level.

Table 4

Amount of data for each targeted acoustic class: speech/non-speech classes, female and male speech classes, telephone and non-telephone speech classes

Corpus	Data amount (s) of each acoustic class					
	Speech	Non-speech	Female	Male	Telephone	non-telephone
<i>RT03S-Dev</i>	3090	321	730	2360	220	2870
<i>RT03S-Eva</i>	4127	478	1271	2856	530	3597

554 tion error goes from 1.5% for the *RT03S-Dev* set at 5.5% for the *RT03S-Eva* set. As said in the
 555 description of the corpora, the reference segmentation provided by NIST does not include tele-
 556 phone/non-telephone information. Therefore, the accuracy of the acoustic segmentation system
 557 for the telephone and non-telephone classification is evaluated using reference boundaries marked
 558 by the authors (*HandS/NS – Gender – T/NT*): less than 0.1% for the *RT03S-Dev* corpus and 3%
 559 for the *RT03S-Eva*.

560 4.3.2. Performance of speaker diarization

561 This section presents the experimental results obtained when applying different levels of acous-
 562 tic macro-class segmentation prior to the speaker diarization systems (*integrated* and *step-by-step*
 563 methods). Experiments are conducted on *ELISA-Dev* and *ELISA-Eva* corpora.

564 Table 5 provides the results obtained individually by each speaker diarization system before
 565 applying the re-segmentation step described in Section 2.4 whereas Table 6 provides the results
 566 obtained after the re-segmentation step. Three kinds of observation may be pointed out through
 567 these results, expressed in terms of missed speaker error rate (*MiE*), false alarm speaker error rate
 568 (*FaE*), speaker error rate (*SE*) and diarization error rate (*DER*):

- 569 • (a) Concerning the corpora (*ELISA-Dev* and *ELISA-Eva*), a large variation in terms of perfor-
 570 mance may be observed between the speaker diarization systems depending on the corpus used.
 571 Indeed, the performance of the *integrated* system drastically decreases on *ELISA-Eva* corpus
 572 compared with *ELISA-Dev* (e.g., from 14.8% to 27.3% for *S/NS-Gender- T/NT* acoustic seg-
 573 mentation) while the *step-by-step* system performance remains quite steady whatever the corpus
 574 used.

Table 5

Error rates, expressed in terms of missed speaker (*MiE*), false alarm speaker (*FaE*), speaker (*SE*) and diarization speaker (*DER*) error rates (%), obtained by each speaker diarization system before applying the re-segmentation step when combined with different levels of acoustic macro-class segmentation

Acoustic segmentation	<i>ELISA-Dev</i>				<i>ELISA-Eva</i>			
	<i>MiE</i>	<i>FaE</i>	<i>SE</i>	<i>DER</i>	<i>MiE</i>	<i>FaE</i>	<i>SE</i>	<i>DER</i>
<i>Step-by-step system</i>								
Hand S/NS-Gender-T/NT	0.0	0.0	14.0	14.0	0.0	0.0	10.2	10.2
S/NS	2.8	2.4	14.5	19.7	2.1	3.0	12.2	17.3
S/NS-Gender	2.8	2.4	13.5	18.7	2.1	3.0	13.6	18.7
S/NS-Gender-T/NT	2.8	2.4	13.9	19.1	2.1	3.0	13.3	18.4
S/NS-Gender-T/S/MS/DS	2.8	2.4	19.5	24.7	2.1	3.0	22.5	27.6
<i>Integrated system</i>								
Hand S/NS-Gender-T/NT	0.0	0.0	10.7	10.7	0.0	0.0	12.0	12.0
S/NS	2.8	2.4	10.2	15.4	2.1	3.0	21.8	26.9
S/NS-Gender	2.8	2.4	9.6	14.8	2.1	3.0	22.2	27.3
S/NS-Gender-T/NT	2.8	2.4	9.9	15.1	2.1	3.0	13.0	18.1
S/NS-Gender-T/S/MS/DS	2.8	2.4	18.0	23.2	2.1	3.0	23.0	28.1

Experiments conducted on *ELISA-Dev* and *ELISA-Eva* corpora.

Table 6

Error rates, expressed in terms of missed speaker (*MiE*), false alarm speaker (*FaE*), speaker (*SE*) and diarization speaker (*DER*) error rates (%), obtained by each speaker diarization system after applying the re-segmentation step when combined with different levels of acoustic macro-class segmentation

Acoustic segmentation	<i>ELISA-Dev</i>				<i>ELISA-Eva</i>			
	<i>MiE</i>	<i>FaE</i>	<i>SE</i>	<i>DER</i>	<i>MiE</i>	<i>FaE</i>	<i>SE</i>	<i>DER</i>
<i>Step-by-step system</i>								
Hand S/NS-Gender-T/NT	0.0	0.0	13.7	13.7	0.0	0.0	10.5	10.5
S/NS	2.8	2.4	13.6	18.8	2.1	3.0	10.3	15.4
S/NS-Gender	2.8	2.4	12.5	17.7	2.1	3.0	10.0	15.1
S/NS-Gender-T/NT	2.8	2.4	12.2	17.4	2.1	3.0	8.6	13.7
S/NS-Gender-T/S/MS/DS	2.8	2.4	12.3	17.5	2.1	3.0	9.4	14.5
<i>Integrated system</i>								
Hand S/NS-Gender-T/NT	0.0	0.0	9.2	9.2	0.0	0.0	10.8	10.8
S/NS	2.8	2.4	10.3	15.5	2.1	3.0	21.4	26.5
S/NS-Gender	2.8	2.4	7.8	13.0	2.1	3.0	19.8	24.9
S/NS-Gender-T/NT	2.8	2.4	7.6	12.8	2.1	3.0	9.0	14.1
S/NS-Gender-T/S/MS/DS	2.8	2.4	9.8	15.0	2.1	3.0	9.0	14.1

Experiments conducted on *ELISA-Dev* and *ELISA-Eva* corpora.

- 575 • (b) The manual acoustic macro-class segmentation gives the best overall DER, although this in
576 part is due to the *MiE* and *FaE* components being zero. Similarly, a large improvement of the
577 *integrated* approach results is obtained with the speech/non-speech, gender and telephone/non-
578 telephone segmentations (*S/NS-Gender-T/NT*), especially on *ELISA-Eva* corpus, without (from
579 26.9% to 18.1%) and with (from 26.5% to 14.1%) the re-segmentation phase. On *ELISA-Dev*
580 corpus, this improvement is more visible after the re-segmentation phase (from 15.5% to
581 12.8%) than before the re-segmentation for which only a small drop is observed (from 15.4%
582 to 15.1%). On the other hand, even if some improvement can be noticed for the *step-by-step*
583 system, the gain is minor. In fact, it is only really visible after the re-segmentation phase (from
584 18.8% to 17.4% on *ELISA-Dev* and from 15.4% to 13.7% on *ELISA-Eva*). Finally, no improve-
585 ment is seen (and in some cases, even a performance loss occurs) when the most detailed acous-
586 tic segmentation (*S/NS-Gender-T/S/MS/DS*) is involved. It can be noticed that this loss of
587 performance becomes quite important without the re-segmentation phase.
- 588 • (c) Applying the re-segmentation step leads to the best performance in most of the cases. This
589 demonstrates its interest while coupled with both speaker diarization strategies.

590 Comparing all the different levels of segmentation granularity (note (b)), the *S/NS-Gender-T/NT*
591 segmentation seems the most helpful for the speaker diarization task, especially for the *integrated*
592 approach. This point is particularly visible for the *ELISA-Eva* corpus for which 20% of speech
593 time (shared among 2 shows over the 3 available in the corpus) is telephone speech against
594 7.7% only for *ELISA-Dev* corpus (mainly present in 1 show over the 6 available).

595 The difference of behaviors in terms of performance (note (b)) between the two speaker diariza-
596 tion systems may be directly linked to the strategies involved for each of them. It seems reasonable
597 that the *step-by-step* approach especially the speaker turn detection step intrinsically behaves as an

598 acoustic class segmentation system, detecting speaker turns as well as acoustic event changes be-
599 fore the clustering phase. In this sense, the a priori acoustic macro-class segmentation becomes
600 useless to improve performance. Obviously this is not true for speech/non-speech detection since
601 the speaker turn detection phase cannot discard non-speech segments automatically without addi-
602 tional processing.

603 Unexpectedly, the most detailed segmentation, *S/NS-Gender-T/S/MS/DS*, does not lead to
604 performance gain and may actually degrade it in most of the cases. This should be due to some
605 speakers present under different acoustic classes (speech over music followed by speech only,
606 classical for news presenters, or in both clean and degraded speech classes depending on the
607 location of interviews for instance) or on the other hand, to the misclassification errors of
608 the acoustic macro-class segmentation system. Since speaker diarization systems are applied
609 independently on each acoustic class, the same speaker may be split under different labels, lead-
610 ing to an increase in speaker error rates. In the same way, increasing the number of acoustic
611 classes creates smaller segments, which may disturb speaker diarization systems. However, these
612 effects are partially overcome thanks to the re-segmentation phase, which may explain that the
613 loss of performance due to *S/NS-Gender-T/S/MS/DS* segmentation is minor after applying the
614 re-segmentation phase.

615 Finally, combining both speaker diarization systems with manual acoustic segmentation out-
616 performs all the automatic ones. However, since the diarization error rate takes both speaker
617 and speech/non-speech error rates into account, the results cannot be compared directly in this
618 case (manual segmentation does not yield any speech/non-speech error rate). Regarding speaker
619 error rate only, the best speaker diarization system (after re-segmentation) based on an automatic
620 acoustic segmentation on *ELISA-Dev* corpus gets 7.6% against 9.2% for manual segmentation. As
621 a result, the speaker diarization system based on an automatic segmentation outperforms (from a
622 pure speaker diarization point of view) the one based on a manual segmentation. In fact, the anal-
623 ysis of the results showed that some segmentation errors, due to some segments falsely split into
624 two different classes (telephone/non-telephone for instance) by the acoustic macro-class segmen-
625 tation system, may be automatically corrected by the re-segmentation step.

626 4.4. NIST-RT03S results

627 This section presents the results obtained during the *NIST-RT03S* evaluation campaign, on the
628 *RT03S-Dev* and *RT03S-Eva* corpora, for both speaker diarization approaches as well as the “fu-
629 sion” systems, described in Sections 3.1 and 3.2. It can be noted that:

- 630 • The “merging” strategy-based system (ELISA1), submitted as ELISA primary system,
631 obtained the second lower diarization error rate compared to the other NIST-RT03S-partici-
632 pant primary systems (NIST).
- 633 • The “hybridization” strategy-based system (ELISA2) (i.e., the CLIPS system followed by the
634 re-segmentation phase), submitted as a secondary ELISA system, outperformed the best pri-
635 mary system and obtained the lowest speaker diarization error rate.

636 Table 7 summarizes the performance achieved by the different systems proposed during the NIST-
637 RT03S. It shows that:

Table 7

Official results obtained for the NIST-RT'03S evaluation campaign

Different error rates (%)	<i>MiE</i>	<i>FaE</i>	<i>SE</i>	<i>DER</i>
CLIPS (CLIPS1)	2.0	2.9	14.3	19.3
CLIPS + re-segment (ELISA2)	1.1	3.8	8.0	12.9
LIA (LIA1)	1.1	3.8	12.0	16.9
LIA variant (LIA2)	1.1	3.8	19.8	24.7
Merging strategy (ELISA1)	1.1	3.8	9.3	14.2

- 638 • Both the CLIPS 1 system *MiE* and *FaE* rates are different from the remaining systems. This is
639 due to the LIA and ELISA system behavior, which work at 0.2 s block level³ (all the segment
640 boundaries are aligned on a 0.2 s scale) whereas the CLIPS system works at a frame level. This
641 gives small differences in the boundary positions of the segments.
- 642 • The LIA E-HMM based primary system (LIA1) improves performance compared with the
643 CLIPS classical approach (CLIPS1) (16.9% *DER* compared to 19.3%). But the second LIA sys-
644 tem (LIA2) based upon a linear model adaptation described in Meignier et al. (2001) obtained
645 only 24.7% *DER*. This difference in terms of performance illustrates the difficulty of adapting a
646 large statistical model in borderline conditions (only few seconds of adaptation data).
- 647 • The *integrated* (E-HMM) method may clearly benefit from some better segment boundaries
648 and longer segments issued from a classical turn detection approach like the CLIPS one, as
649 demonstrated by the large gain of performance (from 16.9% to 12.9% *DER*) reached by the
650 ELISA2 system. Indeed, the re-segmentation phase improves the accuracy of the CLIPS diar-
651 ization and it allows to reduce the diarization error by 33% (relative).
- 652 • The strategy involved in ELISA1 system performs better than ELISA2 over two recordings
653 while a drastic loss is observed on the last recording. The loss on that particular recording is
654 a good example of the limitation of the merging technique explained in Section 3.2 and dis-
655 cussed in the next remark: one of the systems disagreed with the others. This resulted in too
656 many speakers detected and, most important, in the split of a long true speaker into two hyp-
657 othetical speakers, involving a large error rate. Moreover, this observation illustrates the remark
658 made in Section 4.1, in which it is noted that some caution has to be taken when interpreting
659 the results due to the small size of the different corpora. In other words, the generalization
660 problem is apparent here with only three different shows available for testing.
- 661

662 4.4.1. Remarks on the automatic estimate of the number of speakers

663 Concerning the CLIPS system, complementary experiments showed that automatically estimat-
664 ing the number of speakers during the clustering process (as described in Section 2.2.3) generates
665 only about 4% more of absolute diarization error than the optimal number of speakers. The
666 CLIPS algorithm missed only 7% of the real speakers involved in the files (4 speakers out of 57
667 total speakers on the *RT'03S-eva* corpus). It is important to note here that we call “optimal
668 number of speakers” the number of speakers that minimizes the diarization error and not the real

³ See Sections 2.3.2 and 2.4 for details and significance of the 0.3 s block level (for segmentation phase) and the 0.2 s (for re-segmentation).

669 number of speakers involved in the conversation. The optimal number is usually smaller than the
670 real number due to the fact that there are a lot of speakers, especially in broadcast news data, that
671 do not speak enough to train a reliable statistical model (e.g: 4 s during a 30-min file). To illustrate
672 this point, Table 8 presents the speaker diarization error using respectively the optimal, the esti-
673 mated and the real number of speakers obtained on two speech corpora.

674 4.4.2. Remarks on the “fusion” system

675 Concerning the “fusion” strategy, we observed that the label merging method of the four sys-
676 tems generates about 150 potential speakers per show. These speakers correspond generally to:

- 677 • potential speakers that have a large amount of data assigned (>10 s). These speakers can be
678 considered as correct hypothesized speakers;
- 679 • potential speakers generated by few systems, for example the speakers associated with only one
680 short segment (≤ 10 s). These hypothesized speakers can be suppressed (the weight of these
681 speakers on the final scoring is marginal);
- 682 • potential speakers that have a small amount of data scattered between multiple small segments
683 and that can be considered as zones of indecision.

684 We observed also that after the first iteration of the re-segmentation, the number of speakers is
685 already drastically reduced (from 150 to about 50) since speakers associated with indecision zones
686 do not catch any data during the Viterbi decoding and are automatically removed. However, the
687 merging strategy cannot generally solve the wrong behavior of initial systems that split a “true”
688 speaker into two hypothesized speakers, each tied to a long segment. Suppose all systems agree on
689 a long segment except for one which splits this segment into two parts. This would produce two
690 potential speakers (associated with long duration segments) after the merging phase and since we
691 do not do any clustering before re-segmentation, we have generally a “true” speaker split into two
692 hypothesized speakers.

693 5. Ongoing work

694 5.1. Application to other data

695 Though this paper was mainly dedicated to speaker diarization experiments on broadcast news
696 data, our speaker diarization systems were successfully applied to other data types during the last

Table 8
Speaker diarization error rate using different estimates of the number of speakers

Corpus	N_{Opt}	N_{Est}	N_{Real}
<i>RT03S-Dev</i>	14.5	19.7	24.8
<i>RT03S-Eva</i>	14.0	17.1	16.3

The error rate is extracted from the CLIPS NIST-RT03S results. N_{Opt} : optimal number, the number of speakers that gives the lowest DER (46 on *RT03S-Dev*). N_{Est} : estimated number of speakers (47 on *RT03S-Dev*). N_{Real} : real number of speakers (69 on *RT03S-Dev*).

697 NIST evaluations. Table 9 presents a summary of our results obtained since 2002 on different
698 types of data. In 2002, besides the diarization of broadcast news documents, two other tasks were
699 proposed, namely on telephone speech conversations and on meeting room recordings. Perform-
700 mance shown on the first line of the table illustrates the increasing difficulty of the tasks. For tele-
701 phone conversations, only two people are involved and there are few overlapping segments. For
702 broadcast news, there are obviously more speakers on the audio documents, but this is mostly pre-
703 pared speech with a large part of “studio quality” voice. The hardest task definitely corresponds
704 to meeting data with very spontaneous speech, overlapping voices, disfluencies, distant speakers
705 (in case of table microphones) and background noise. The second line shows the best performance
706 obtained in spring 2003 on broadcast news data with the system described in this paper, and illus-
707 trates the progress made from 2002 to spring 2003 on this data. During the NIST 2004 spring rich
708 transcription evaluation (Fredouille et al., 2004), the novelty was that we had to process multiple
709 speech channels coming from multiple sensors located in a smart meeting room. We proposed a
710 very straightforward strategy to merge the multiple channel segmentation outputs and obtained
711 the best speaker diarization performance for this task (Fredouille et al., 2004).

712 5.2. Integration of a priori knowledge

713 One of the main assumptions in most of the papers (Adami et al., 2002; Meignier et al., 2001)
714 concerning the speaker diarization task is that no a priori information is available on the test data.
715 This means for instance, that the number and the identity of the speakers involved in conversation
716 are not known. A consequence of this hypothesis is that no reference speaker data are supposed to
717 be available before segmenting an audio signal for instance.

718 However, this limitation may not be so necessary (Moraru et al., 2004) for some applications
719 and conditions for which we can reasonably hope to have a priori information. For instance, the
720 type of conversation is generally known (broadcast news, telephone or meeting conversation),
721 which gives us information on speech quality and average speaker turn length. We may also know
722 about the real number of speakers involved in conversation. For instance, we know there are two
723 speakers in telephone conversation while a list of participants might be available for meeting data.
724 In some cases, reference data might be available for the speakers involved in the conversation. For
725 example, we could ask every participant to introduce himself at the beginning of meetings. A list
726 of what kind of information we might expect for each type of audio document is presented in
727 Table 10.

728 Some results concerning the knowledge of the real number of speakers were already presented
729 in Section 4.4. From those results only it was difficult to conclude if the knowledge of the number
730 of speakers is useful information or not since the conclusion is different for the two speech
731 corpora.

Table 9
ELISA Results since 2002 (diarization error rate) given various corpora

Corpus/year	Telephone	Broadcast news	Meeting (head microphone)	Meeting (table microphone)
2002	5.7	30.3	34.7	36.9
Spring 2003	×	12.9	×	×
Spring 2004	×	×	×	22.4

Table 10

Available a priori information for each audio document type

Information	Telephone	Meeting	Broadcast news
No. of speakers	2	Possibly known	Unknown
Reference data available	Possibly (2 speakers)	Possibly (1:N speaker)	Possibly (few speaker only)

732 In the case of broadcast news data, we can easily obtain reference data of one particular speak-
733 er. This speaker is the news show presenter. From previously broadcasted shows, it is thus pos-
734 sible to obtain enough data to train a presenter speaker model directly by EM. We have shown
735 in Moraru et al. (2004) that using a simple speaker tracking system for this particular speaker,
736 up to 3% absolute diarization error reduction can be obtained (experiments done on the ESTER⁴
737 radio broadcast corpus, see Moraru et al., 2004, for more experimental details).

738 The possibility of having reference data for all speakers is specific to telephone and meeting
739 data. The main interest in having data available for all speakers, other than a significant error
740 reduction (up to 10% absolute for the *RT03-Eva* corpus, see Moraru et al., 2004, for experimental
741 details) is the execution time consideration. Our *step-by-step* speaker diarization approach takes
742 about four times realtime for a 30 min file. When reference data are available for all speakers, the
743 speaker diarization is a simple assignment of every segment to the most likely speaker. In this case
744 the speaker diarization takes about 0.1 times realtime. Every speaker model is derived from a
745 background model using the few seconds available for each speaker.

746 In conclusion, with the current modeling technique used in our systems, the real number of
747 speakers is however useful to fix a starting number of classes before the speaker clustering process
748 whereas reference data for the speakers involved in conversation is always useful. Further exper-
749 iments should be done on telephone and meeting data concerning the case when reference data are
750 available for all speakers.

751 6. Conclusion

752 This paper summarizes the collaboration of the LIA and CLIPS laboratories, two members of
753 the ELISA consortium, in the area of speaker diarization. The work presented in this paper was
754 done in the framework of the NIST Rich Transcription 2003 spring evaluation campaign (NIST-
755 RT'03S) and addressed two main points.

756 Firstly, two main approaches for speaker diarization were proposed and compared. The first
757 one relies on a classical strategy based on speaker-turn detection followed by a clustering process,
758 while the second one relies on an integrated strategy where segment boundaries and speaker tying
759 of the segments are extracted simultaneously and challenged during the whole process. The *inte-*
760 *grated* method (E-HMM) shows a higher modeling power (16.9% diarization error compared to
761 19.3% for the *step-by-step* approach). Nevertheless, the classical *step-by-step* approach seems to
762 obtain more consistent results across the files and conditions than the *integrated* one. Despite
763 the differences between the two approaches, the results obtained during the NIST 2003 spring

⁴ www.afcp-parole.org/ester/.

764 evaluation showed the interest of using both techniques. This was confirmed by the results ob-
765 tained using a fusion of both systems where the *integrated* approach is applied after the CLIPS
766 *step-by-step* segmentation system. The “fusion” system obtained a speaker diarization error rate
767 of 12.9% against 19.3% for the CLIPS system used on its own. The “fusion” system also showed a
768 relative 33% error-reduction compared to the performance of the *integrated* system taken alone
769 (from 16.9% to 12.9%). The *integrated* (E-HMM) method clearly benefits from the better segment
770 boundaries and longer segments issued from the classical CLIPS approach. This “fusion” system
771 achieved the lowest speaker diarization error rate during the NIST-RT’03S evaluation campaign.
772 More investigation is needed for a better understanding of the nature of the errors made by the
773 systems, which is not a trivial task as the speaker diarization performance metric is complicated.
774 The second main issue addressed in this work concerns the nature of the audio documents to
775 segment. This paper focuses on the case of audio broadcast news documents. An acoustic macro-
776 class segmentation was proposed, as a prior step for speaker diarization systems. The speaker
777 diarization system is run independently on each acoustic sub-class and the resulting segmentations
778 are merged thanks to a re-segmentation algorithm (in this paper, the re-segmentation process con-
779 sists of one iteration of the *integrated* E-HMM algorithm). For a speech/non-speech, gender and
780 bandwidth (studio/telephone speech) acoustic segmentation, a significant gain was observed in the
781 case of the *integrated* approach. A slight gain was also observed for the *step-by-step* approach,
782 which seems more robust to channel or environment variations. Moreover, finer macro-class seg-
783 mentation (including speech over music detection) led to a loss in performance, partially due to
784 the assumption (fixed during the speaker segmentation process) that the same speaker could
785 not appear in more than one acoustic macro-class.

786 Finally, some ongoing work has been presented, which demonstrates that the approaches pro-
787 posed are able to deal with new types of data like meeting room recordings. Indeed, multiple
788 microphones are often available in the case of meeting rooms and taking these multiple (and
789 low quality) recordings of the same conversation into account constitutes a new challenge for
790 speaker diarization. The systems presented in this paper were adapted to the meeting task and ob-
791 tained during the NIST-RT’04S campaign a state-of-the-art result with 22.8% diarization error
792 rate. A strategy for fusing segmentations issued from several microphones was also proposed
793 but no significant win was observed compared with results obtained using the best microphone.
794 Secondly, the interest of a priori knowledge concerning the potential number of speakers or
795 the speakers themselves was also presented. In particular, using knowledge from well-known
796 speakers allows a 3% absolute win during experiments on the ESTER database and knowing
797 all the potential speakers allows us to speed up very significantly the diarization process. This pre-
798 liminary work on using available a priori information opens up interesting possibilities for further
799 work.

800 References

- 801 Adami, A., Kajarekar, S.S., Hermansky, H., 2002. A new speaker change detection method for two-speaker
802 segmentation. In: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP
803 2002), vol. IV, pp. 3908–3911.
- 804 Ajmera, J., Wooters, C., 2003. A robust speaker clustering algorithm. In: Automatic Speech Recognition and
805 Understanding, IEEE, ASRU 2003, St. Thomas, US Virgin Islands, pp. 411–416.

- 806 Chen, S., Gopalakrishnan, P., 1998. Speaker, environment and channel change detection and clustering via the bayesian
807 information criterion. In: DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne,
808 VA.
- 809 Darpa speech recognition evaluation workshop. Available from: <<http://www.nist.gov/speech/publications/>>.
- 810 Delacourt, P., Welkens, C.J., 2000. DISTBIC: a speaker based segmentation for audio data indexing. *Speech*
811 *Communication* 32, 111–126.
- 812 ELISA, 2000. The ELISA systems for the NIST 99 evaluation in speaker detection and tracking. *Digital Signal*
813 *Processing (DSP), a review journal – Special issue on NIST 1999 speaker recognition workshop 10 (1–3)*, pp. 143–
814 153.
- 815 Fredouille, C., Moraru, D., Meignier, S., Besacier, L., Bonastre, J.-F., 2004. The NIST 2004 spring rich transcription
816 evaluation: two-axis merging strategy in the context of multiple distance microphone based meeting speaker
817 segmentation. In: *RT2004 Spring Meeting Recognition Workshop*, p. 5.
- 818 Gauvain, J.-L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of
819 Markov chains. *IEEE Transactions on Speech and Audio Processing* 22, 291–298.
- 820 Gauvain, J.-L., Lamel, L., Adda, G., 1998. Partitioning and transcription of broadcast news data, In: *Proceedings of*
821 *International Conference on Spoken Language Processing (ICSLP 98)*.
- 822 Gauvain, J.-L., Lamel, L., Adda, G., 2001. Audio partitioning and transcription for broadcast data indexation.
823 *Multimedia Tools and Applications*, 187–200.
- 824 Gauvain, J.-L., Lamel, L., Adda, G., 2002. The LIMSI broadcast news transcription system. *Speech Communication* 37
825 (1–2), 89–108.
- 826 Hain, T., Woodland, P., 1998. Segmentation and classification of broadcast news audio. In: *Proceedings of*
827 *International Conference on Spoken Language Processing (ICSLP 98)*, Sydney, Australia.
- 828 Kim, D.Y., Evermann, G., Hain, T., Mrva, D., Tranter, S., Wang, L., Woodland, P.C., 2003. Recent advances in
829 broadcast news transcription. In: *Automatic Speech Recognition and Understanding, IEEE, ASRU 2003*, St.
830 Thomas, US Virgin Islands, pp. 105–110.
- 831 Magrin-Chagnolleau, I., Gravier, G., Blouet, R., 2001. for the ELISA consortium, Overview of the ELISA consortium
832 research activities. In: *2001: A Speaker Odyssey. The Speaker Recognition Workshop, Chania, Crete*, pp. 67–72.
- 833 Meignier, S., Bonastre, J.-F., Fredouille, C., Merlin, T., 2000. Evolutive HMM for speaker tracking system. In:
834 *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2000)*, Istanbul,
835 Turkey, pp. 1177–1180.
- 836 Meignier, S., Bonastre, J.-F., Igounet, S., 2001. E-HMM approach for learning and adapting sound models for speaker
837 indexing. In: *2001: a Speaker Odyssey. The Speaker Recognition Workshop, Chania, Crete*, pp. 175–180.
- 838 Moraru, D., Meignier, S., Besacier, L., Bonastre, J.-F., Magrin-Chagnolleau, Y., 2003. The ELISA consortium
839 approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation. In: *Proceedings of*
840 *International Conference on Acoustics Speech and Signal Processing (ICASSP 2003)*, vol. II, Hong Kong, pp. 89–
841 92.
- 842 Moraru, D., Meignier, S., Fredouille, C., Besacier, L., Bonastre, J.-F., 2004. The ELISA consortium approaches in
843 broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. In: *Proceedings of*
844 *International Conference on Acoustics Speech and Signal Processing (ICASSP 2004)*, Montreal, Canada.
- 845 Moraru, D., Besacier, L., Castelli, E., 2004. Using a priori information for speaker diarization. In: *2004: A Speaker*
846 *Odyssey. The Speaker Recognition Workshop, Toledo, Spain*, pp. 355–362.
- 847 Nguyen, L., Xiang, B., 2004. Light supervision in acoustic model training. In: *Proceedings of International Conference*
848 *on Acoustics Speech and Signal Processing (ICASSP 2004)*, Montreal, Canada.
- 849 NIST, Reference data cookbook for who spoke when diarization task. Available from: <[http://www.nist.gov/speech/](http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/ref-cookbook-v2_4.pdf)
850 [tests/rt/rt2003/spring/docs/ref-cookbook-v2_4.pdf](http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/ref-cookbook-v2_4.pdf)>, v2.4 (2003).
- 851 NIST, Rt-03s workshop agenda and presentations. Available from: <[http://www.nist.gov/speech/tests/rt/rt2003/](http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations)
852 [spring/presentations](http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations)>.
- 853 NIST, The NIST 2001 speaker recognition evaluation plan. Available from: <[http://www.nist.gov/speech/tests/spk/](http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrac-evalplan-v05.9.pdf)
854 [2001/doc/2001-spkrac-evalplan-v05.9.pdf](http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrac-evalplan-v05.9.pdf)> (March 2001).
- 855 NIST, The NIST year 2002 speaker recognition evaluation plan. Available from: <[http://www.nist.gov/speech/tests/](http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrac-evalplan-v60.pdf)
856 [spk/2002/doc/2002-spkrac-evalplan-v60.pdf](http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrac-evalplan-v60.pdf)> (February 2002).

- 857 NIST, The rich transcription spring 2003 (RT-03S) evaluation plan. Available from: <<http://www.nist.gov/speech/858 tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf>>, (Version 4, Updated 02/25/2003) (February 2003).
- 859 NIST, Spring 2004 (rt-04s) rich transcription meeting recognition evaluation plan. Available from: <<http://860 www.nist.gov/speech/tests/rt/rt2004/spring/documents/rt04s-meeting-eval-plan-v1.pdf>> (February 2004).
- 861 Quénot, G., Moraru, D., Besacier, L., Mulhem, P., 2002. Clips-imag at trec-11: Experiments in video retrieval. In:
862 TREC 2002, Gaithersburg, MD, USA.
- 863 Quénot, G., Moraru, D., Besacier, L., 2003. Clips at trecvid: Shot boundary detection and feature detection. In: TREC
864 2003, Gaithersburg, MD, USA.
- 865 Reynolds, D.A., Dunn, R.B., Laughlin, J.J., 2000. The Lincoln speaker recognition system: NIST EVAL2000. In:
866 Proceedings of International Conference on Spoken Language Processing (ICSLP 2000), vol. 2, Beijing, China, pp.
867 470–473.
- 868 Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models, Digital
869 Signal Processing (DSP), a review journal – Special issue on NIST 1999 speaker recognition workshop 10 (1–3), pp.
870 19–41.
- 871 Schwarz, G., 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2), 461–464.
- 872 Siegler, M., Jain, U., Raj, B., Stern, R., 1997. Automatic segmentation and clustering of broadcast news audio. In: the
873 DARPA Speech Recognition Workshop, Westfields, Chantilly, Virginia.
- 874 Siu, M.-H., Rohlicek, R., Gish, H., 1992. An unsupervised, sequential learning algorithm for segmentation of speech
875 waveforms with multi-speakers. In: Proceedings of International Conference on Acoustics Speech and Signal
876 Processing (ICASSP 92), vol. 2, San Francisco, CA, pp. 189–192.
- 877 Smeaton, A., Kraaij, W., Over, P., 2003. TRECVID 2003 – an introduction. In: 12th Text Retrieval Conference.
- 878 Wilcox, L., Chen, F., Kimber, D., Balasubramanian, V., 1994. Segmentation of speech using speaker identification, In:
879 Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 94), Adelaide,
880 Australia, pp. 161–164.
- 881 Wilcox, L., Kimber, D., Chen, F., 1994. Audio indexing using speaker identification. In: Proceedings SPIE Conference
882 on Automatic Systems for the Inspection and Identification of Humans, San Diego, CA, pp. 149–157.
- 883 Woodland, P., 2002. The development of the HTK broadcast news transcription system: an overview. *Speech*
884 *Communication* 37 (1–2), 291–299.
- 885