Multistage Speaker Diarization of Broadcast News

Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain, *Member, IEEE*

***Abstract*—This paper describes recent advances in speaker diarization with a multistage segmentation and clustering system, which incorporates a speaker identification step. This system builds upon the baseline audio partitioner used in the LIMSI broadcast news transcription system. The baseline partitioner provides a high cluster purity, but has a tendency to split data from speakers with a large quantity of data into several segment clusters. Several improvements to the baseline system have been made. First, the iterative Gaussian mixture model (GMM) clustering has been replaced by a Bayesian information criterion (BIC) agglomerative clustering. Second, an additional clustering stage has been added, using a GMM-based speaker identification method. Finally, a post-processing stage refines the segment boundaries using the output of a transcription system. On the National Institute of Standards and Technology (NIST) RT-04F and ESTER evaluation data, the multistage system reduces the speaker error by over 70% relative to the baseline system, and gives between 40% and 50% reduction relative to a single-stage BIC clustering system.**

***Index Terms*—Bayesian information criterion (BIC) clustering, speaker diarization, speaker identification (SID), speaker segmentation and clustering.**

I. Introduction

SPEAKER diarization, also called speaker segmentation and clustering, is the process of partitioning an input audio stream into homogeneous segments according to speaker identity. It is one aspect of audio diarization, along with categorization of music, background noise, and channel conditions. Speaker diarization can improve the readability of an automatic transcription by structuring the audio stream into speaker turns and in some cases by providing the true speaker identity. Such information can also be of interest for the indexing of multimedia documents. As defined by the National Institute of Standards and Technology (NIST) for the 2004 Rich Transcription evaluation [1], the speaker diarization task is relative to a given show and no *a priori* knowledge of the speakers voice or even of the number of speakers is available. Therefore, only a relative, show-internal speaker identification is produced by the diarization system. This definition has been adopted in this paper, even though a speaker diarization system could obviously make use of such information if combined with a speaker identification or tracking system.

Audio diarization is a useful preprocessing step for an automatic speech transcription system. By separating out speech and nonspeech segments, the recognizer only needs to process audio segments containing speech, thus reducing the computation time and avoiding word insertions in these portions. By clustering segments of the same acoustic nature, condition-specific models can be used to improve the quality of the transcription. By clustering segments from the same speaker, the amount of data available for unsupervised speaker adaptation is increased, which can significantly improve the transcription performance.

Automatic speech transcription and speaker diarization rely on similar methods for segmentation and clustering. However, differences in their objectives leads to different needs, particularly concerning where accuracy is most important. Automatic transcription requires accurate segment boundaries. Although the rejection of nonspeech segments is useful in order to minimize insertion of words and to save computation time, it is important that the segment boundaries are located in noninformative zones such as silences or breaths. Indeed, having a word cut by a boundary disturbs the transcription process and increases the word error rate.

Diarization also aims to produce homogeneous speech segments; however, the main objectives are the purity and the correct labeling of the segments. Errors such as having more than one cluster for a given speaker, or conversely, merging the segments of two different speakers into one cluster, are penalized more heavily. The effects of both boundary inaccuracy and mislabeled segments were measured in [2] for English broadcast news. The experiments showed that segment boundary errors have a greater impact on the transcription task, while label errors have a greater impact on the diarization task.

The NIST Rich Transcription evaluation has been the major evaluation for speaker diarization of broadcast news data in 2003 and 2004 [1], [3], [4]. In 2005, the Technolangue ESTER evaluation was conducted on a similar task using French radio broadcast news data [5], [6].

Most speaker diarization systems for broadcast news data have a similar general architecture. First, the signal is chopped into homogeneous segments. The segment boundaries are located by finding acoustic changes in the signal and each segment is expected to contain speech from only one speaker. The resulting segments are then clustered so that each cluster corresponds to one speaker, a major issue being that the number of speakers is unknown *a priori* and needs to be automatically determined. Each system also presents specific aspects which can be classified following different criteria.

Manuscript received October 15, 2005; revised February 15, 2006. This work was supported in part by the European Commission under the FP6 Integrated Project IP 506909 CHIL and in part by the Defense Advanced Research Projects Agency under the GALE Program. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA. The associate editor coordinating the review of this paper and approving it for publication was Dr. Geoffrey Zweig.

C. Barras, X. Zhu, and J.-L. Gauvain are with the Computer Sciences Laboratory for Mechanics and Engineering Sciences-National Center for Scientific Research (LIMSI-CNRS), 91403 Orsay, France (e-mail: barras@limsi.fr; xuan@limsi.fr; gauvain@limsi.fr).

S. Meignier was with the Computer Sciences Laboratory for Mechanics and Engineering Sciences-National Center for Scientific Research (LIMSI-CNRS), 91403 Orsay, France. He is now with the Laboratoire d'Informatique de l'Universite du Maine, 72085 Le Mans Cedex 9, France (e-mail: sylvain.meignier@lium.univ-lemans.fr).

Digital Object Identifier 10.1109/TASL.2006.878261

- *Link between segmentation and clustering:* Segmentation can be done first, followed by clustering with no connection between the two parts [7], [8] inspired from the work presented in [9]–[11]; alternatively, the segmentation and clustering can be jointly optimized, via, for example, the iterative segmentation and clustering procedures described in [12]–[14]. A limitation of the first method is that errors made in the segmentation step are not only difficult to correct later, but can also degrade the performance of the subsequent clustering step.
- *Clustering strategy:* It relies either on an agglomerative clustering [7], [12], [14] or on a divisive clustering method [13], [15].
- *Modeling strategy:* Each speaker can be modeled by a Gaussian mixture model (GMM) with diagonal covariance matrices composed of eight to 64 components. As is done in the speaker recognition task, larger models with 2048 components have been proposed [8], [13], [14]. In this case, a more robust estimation of the models despite the limited amount of data per speaker can be obtained by performing the maximum *a posteriori* (MAP) adaptation of a prior model [16]. On the other hand, using a single Gaussian with a full covariance matrix for the modeling of a speaker also provides good results [7].

In our experiments, several variants and combinations of systems have been tested, in particular to study the link between segmentation and clustering and the modeling strategy.

The remainder of this paper is organized as follows. Section II describes the baseline partitioning system which was developed for the automatic broadcast news transcription task. Section III describes the multistage partitioning system specifically built for the speaker diarization task. This system is based upon a Bayesian information criterion (BIC) clustering followed by speaker identification (SID) clustering. Experimental results are presented in Section IV, followed by some conclusions.

II. Baseline Partitioning System

The baseline audio partitioning system was developed as a preprocessing step for the LIMSI English broadcast news transcription system [12], [17]. It was shown to provide a high cluster purity (about 96%) and a cluster coverage slightly below 80% on 1996 and 1997 NIST evaluation data. This baseline partitioner **c-std** is shown in Fig. 1.

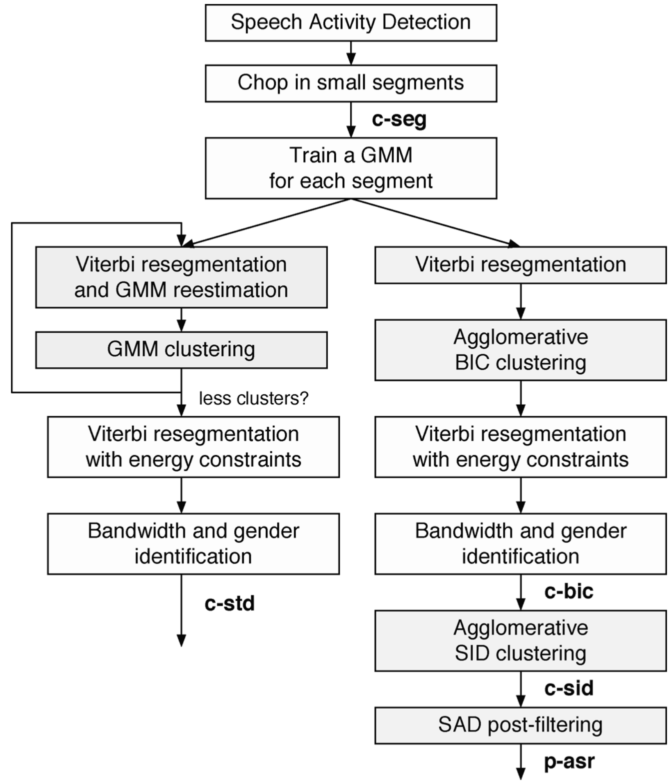

Fig. 1. Architecture of the baseline partitioning system (**c-std** on the left side of the diagram) and the multistage speaker diarization system (**p-asr** to the right, along with **c-seg**, **c-bic**, and **c-sid** intermediate steps).

A. Feature Extraction

Mel frequency cepstral parameters are extracted from the speech signal every 10 ms using a 30-ms window on a 0–8 kHz band. For each frame, the Mel scale power spectrum is computed and the cubic root taken followed by an inverse Fourier transform. Using a process similar to that of Perceptual linear predictive (PLP) computation [19], 12 LPC-based cepstral coefficients are then extracted. The 38-dimensional feature vector consists of 12 cepstral coefficients, Δ and $\Delta - \Delta$ coefficients plus the Δ and $\Delta - \Delta$ log-energy. This is essentially the same set of features that is used in a standard transcription system, except for the energy [18]. This set is used in all steps of the **c-std** system, except for the segmentation into small segments where only the static features are used. No cepstral mean or variance normalization is performed to the acoustic vector in the baseline partitioning system.

B. Speech Activity Detection (SAD)

Speech is extracted from the signal with a Viterbi decoding using GMM for speech, noisy speech, speech over music, pure music, and silence or noise. The aim of the SAD is to remove only long regions without speech such as silence, music, and noise, so the penalty of switching between models in the Viterbi decoding was set to minimize the loss of speech signal. The GMMs, each with 64 Gaussians, were trained on about 1 h of the specific type of data, selected from English broadcast news data from 1996 and 1997 distributed by the Linguistic Data Consortium (LDC).

C. Chopping Into Small Segments

The segmentation process consists of finding segment boundaries that correspond to the instantaneous speaker change points. It is generally a good choice to minimize the miss rate for speaker change points even if the false alarm rate is high, because the false change points can be easily removed later during a clustering process. The segmentation needs to provide pure segments (i.e., containing speech from only one speaker) of duration sufficient to characterize the voice of the speaker.

Segmentation of the signal is performed by taking the maxima of a local Gaussian divergence measure between two adjacent sliding windows s_1 and s_2. For each segment, the static features (i.e., only the 12 cepstral coefficients plus the energy) are modeled with a single diagonal Gaussian, i.e., $s_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $s_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ with Σ_1 and Σ_2 diagonal. Then the Gaussian divergence measure is defined as

$$G(s_1, s_2) = (\mu_2 - \mu_1)' \Sigma_1^{-1/2} \Sigma_2^{-1/2} (\mu_2 - \mu_1). \quad (1)$$

It is the Mahalanobis distance between μ_1 and μ_2 weighted by the geometric mean of Σ_1 and Σ_2, which reduces to a weighted Euclidean distance because of the diagonal assumption. The detection threshold was optimized on the training data in order to provide acoustically homogeneous segments. The window size was set to 5 s with a minimal segment length of 2.5 s. Due to the simple diagonal assumption, this segmentation phase is very quick. This approach is similar to the segmentation proposed in [9] using the symmetric KL2 metric. Other popular segmentation methods are based upon the BIC metric [10], [20], [21], but these methods show a much higher complexity [22]. An analysis of various speaker change point techniques based on models, metrics, or energy is given in [23].

D. Iterative GMM Segmentation/Clustering Procedure

Each initial segment is used to seed one cluster, and an eight-component GMM with a diagonal covariance matrix is trained on the segments data. The algorithm alternates the Viterbi resegmentation and the GMM re-estimation and merging steps, with the goal of maximizing the objective function

$$\sum_{i=1}^{N} \log f(s_i | M_{c_i}) - \alpha N - \beta K \quad (2)$$

where $S = (s_1, \ldots, s_N)$ is the partitioning of the speech segments into a sequence of N segments, $c_i \in [1, K]$ is the cluster label for the segment s_i among the K different clusters, $f(s_i | M_{c_i})$ is the likelihood of the segment s_i given the model of its cluster M_{c_i}, and α and β are the segment and cluster penalties. The procedure stops when no more merges are possible. More details on the clustering procedure can be found in [12]. This procedure is similar to BIC using a global penalty as described in Section III-A.

E. Viterbi Resegmentation

The segment boundaries are refined using the last set of GMMs and an additional relative energy-based boundary penalty, within a 1-s interval. The boundaries are thus shifted to the nearest point of low energy within this interval. This is done so as to locate the segment boundaries at silence portions, thereby avoiding cutting words. This is especially important when using the resulting segmentation as a preprocessing step of an automatic transcription system.

F. Bandwidth and Gender Labeling

Bandwidth (wide-band studio or narrow-band telephone) detection for each segment is first performed using two GMMs. The gender (male or female) labeling is then carried out on the segments using two pairs of bandwidth-dependent GMMs: for gender labeling on telephone speech segments, feature extraction is limited to the 0–3.5 kHz band. The GMM models are composed of 64 components with diagonal covariance matrices and were trained on the subset of the LDC 1996/1997 English broadcast news data also used to train the speech detection models. This labeling is useful for the transcription system, as different acoustic models are used for each combination of bandwidth and gender for better performance, but is also of interest for structuring the acoustic stream. Performing the labeling on a segment basis rather than for a whole cluster may split a cluster in two which can prove beneficial, since a given speaker is usually not recorded in both wide-band and narrow-band conditions in the same show.

III. Multistage Diarization

Recent research has shown BIC clustering methods to obtain good performance on the speaker diarization task [7], [15]. We therefore tested a modified system, replacing the iterative GMM clustering with a BIC-based clustering (cf. Fig. 1) and using Gaussian models with a full covariance matrix. We believe the iterative resegmentation to be more important in the context of speaker partitioning for the transcription task than for the speaker diarization task. Since different models can capture different and complementary aspects of the data, we decided to combine them in a multistage system. We decided to pipeline the output of the system with the BIC clustering into a second clustering stage which uses a speaker identification module. The SID clustering uses a more aggressive acoustic channel normalization and a more complex speaker model, enabled by the larger amount of data per cluster at the beginning of this stage. Finally, a SAD post-filtering stage was added in order to remove short intraspeaker pauses. These short pauses while indeed harmless for a speech transcription system, are penalized as false alarms in a speaker diarization system. The other parts of the system were kept unchanged.

A. BIC Clustering

Agglomerative clustering is applied to the segments resulting from the GMM segmentation. Initially, each segment seeds one cluster, modeled by a single Gaussian with a full covariance matrix trained on the 12 Mel frequency cepstrum coefficients and the energy (but without the Δ coefficients). At each iteration, the two nearest clusters are merged until the stopping criterion is reached. The BIC criterion [10] is used both for the intercluster distance measure and the stop criterion.

In order to decide whether to merge two clusters c_i and c_j, the ΔBIC value is computed as

$$\Delta\mathrm{BIC} = (n_i + n_j) \log|\Sigma| - n_i \log|\Sigma_i| - n_j \log|\Sigma_j| - \lambda P \quad (3)$$

where Σ is the covariance matrix of the merged cluster (c_i and c_j), Σ_i of cluster c_i, Σ_j of cluster c_j, and n_i and n_j are, respectively, the number of the acoustic frames in clusters c_i and c_j. The penalty P is

$$P = \frac{1}{2}\left(d + \frac{1}{2} d(d+1)\right) \log n \quad (4)$$

where d is the dimension of the feature vector space. The term $n_i \log|\Sigma_i|$ is related to the log likelihood of the cluster c_i given its estimated Gaussian M_{c_i}.[1] Singular covariance matrices were not an issue because of the minimal length constraint during the initial segmentation. The merging criterion is that two clusters should be merged if $\Delta\mathrm{BIC} < 0$. At each step the two nearest clusters (i.e., those which have the most negative $\Delta\mathrm{BIC}$ value) are merged into one cluster, and the $\Delta\mathrm{BIC}$ values between this new cluster and remaining clusters are computed. This clustering procedure terminates when the $\Delta\mathrm{BIC}$ between all cluster pairs is greater than zero.

In our BIC clustering procedure, the size of the two merged clusters, i.e., $n = n_i + n_j$, is used to compute the penalty P, as described in [21]. We refer to this as a local BIC penalty. Another solution is to use the size of the whole set of clusters, i.e., $n = \sum_{k=1}^{N} n_k$ to compute the penalty, which we refer to as a global BIC penalty and corresponds to an exponential prior for the number of clusters. In this case, the penalty is constant and the decision to merge two clusters is decided just by the increase in likelihood. This in fact corresponds to the objective function in (2) used in the baseline partitioner when the number of segments is fixed. For broadcast news documents, our experimental results as presented in Section IV demonstrate the local BIC to be a better choice for a merging criterion.

B. SID Clustering

After the initial segmentation, both the iterative GMM and the agglomerative BIC clustering methods have to deal in the beginning of the process with short-duration segments, and thus use a limited set of parameters per cluster. After several iterations, the amount of data per cluster increases, so a more complex model can be used. Our approach is to stop the initial clustering stage early, and use the results to seed a second clustering stage with more initial data per cluster. This second stage can therefore estimate more complex models for the speakers. In addition, purely acoustic clustering tends to split a speakers data into several clusters as a function of the various background conditions (clean speech, speech with noise, speech with music, etc.), so an acoustic background normalization is necessary to regroup the data for a given speaker.

After the BIC clustering stage, state-of-the-art speaker recognition methods [24], [25] were used to improve the quality of the speaker clustering. The feature vector consists of 15 Mel frequency cepstral coefficients plus delta coefficients and delta energy. Feature warping normalization, which reshapes the histogram of the cepstral coefficients into a Gaussian distribution [26] is performed on each segment using a sliding window of 3 s in order to reduce the effect of the acoustic environment.

For each gender and channel condition (studio, telephone) combination, a universal background model (UBM [27]) with 128 diagonal Gaussians is trained on the 1996/1997 English broadcast news data. The GMM for each remaining cluster is obtained by MAP adaptation [16] of the means of the matching UBM.

Agglomerative clustering is performed separately for each gender and bandwidth condition, using a cross log-likelihood ratio as in [28]. For each cluster c_i, its model M_i is MAP adapted from the gender and channel-matched UBM B using the feature vectors x_i belonging to the cluster. Given two clusters c_i and c_j, the cross log-likelihood ratio $\mathcal{S}$ is defined as

$$\mathcal{S}(c_i, c_j) = \frac{1}{n_i}\log\frac{f(x_i|M_j)}{f(x_i|B)} + \frac{1}{n_j}\log\frac{f(x_j|M_i)}{f(x_j|B)} \quad (5)$$

where $f(\cdot|M)$ is the likelihood of the acoustic frames given the model M, and n_i is the number of frames in cluster c_i. This is a symmetric similarity measure. After each merge, a new model is trained for the cluster $c_i \cup c_j$. The clustering stops when the cross log-likelihood ratio between all clusters is below a given threshold δ optimized on the development data.

C. SAD Post-Filtering

In order to filter out short-duration silence segments that were not removed in the initial speech detection step to further reduce the speaker diarization error, a post-processing stage uses the word segmentation output by the LIMSI Broadcast News speech-to-text system [29] relying on the **c-std** system for the segmentation and clustering. Only interword silences shorter than 1 s are filtered out, this value being determined empirically.

IV. Experiments and Results

Experiments are reported for diarization systems that were submitted to the NIST and ESTER evaluations [30]. Several configurations were tested for the systems. Unless otherwise specified, the configuration used is the one that provided the best results on development data, i.e., $\alpha = \beta = 230$ for **c-std**, $\lambda = 5.5$ for **c-bic**, and $\lambda = 3.5$, $\delta = 0.1$ for **c-sid** and **p-asr**. In the **c-sid** system, the BIC penalty weight λ was optimized to cluster only the closest segments in the BIC clustering stage so as to give more degrees of freedom to the SID clustering stage; while in the **c-bic** system, λ was optimized directly to give the lowest diarization error the BIC clustering could bring. A local BIC merging and stop criterion was also used.

A. Corpora

The experiments were conducted on the U.S. English data used in NIST RT-04F (Fall 2004 Rich Transcription evaluation) [1] and on the French data from the French ESTER broadcast news evaluation [5].

For the RT-04F evaluation, the training and development corpora were provided for system development along with a reference speaker labeling determined by LDC. Evaluation references were made available after the evaluation. The data are from U.S. radio or TV broadcast news shows. The development data has two portions, "*dev1*" and "*dev2*" each consisting of six 30-min audio files. *Dev1* was recorded in February 2001, with programs from ABC, CNN, NBC, PRI, and VOA; and *dev2* was recorded in November and December 2003, with programs from ABC, CNBC, CNN, CSPAN, and PBS. The evaluation data is comprised of 12 audio files each lasting approximately 30 min, recorded in December 2003, extracted from the shows of ABC, CNBC, CNN, CSPAN, PBS, and WBN. The *dev2* and evaluation and corpora are very similar to each other since the

[1]More precisely, $\log f(c_i|M_{c_i}) = -(n_i/2)\log|\Sigma_i| - (n_i d/2)(1 + \log 2\pi)$, but the constant factor 1/2 was simplified in (3).

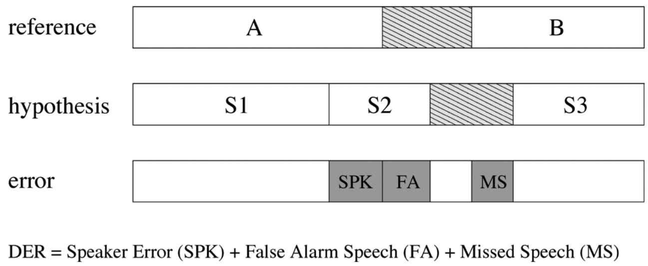

Fig. 2. Example of the performance measures for the speaker diarization task used in the NIST RT-04F evaluation.

shows are recorded from the same channels, at the same period, whereas *dev1* is an older corpus coming from the previous RT-03 evaluation.

The data used in ESTER were extracted from French radio broadcast news shows, provided by the Evaluations and Language resources Distribution Agency (ELDA) and the Délégation Générale pour l'Armement (DGA). The training corpus contains 82 h of data from the France Inter, France Info, RFI, and RTM radio stations, recorded in 1998, 2000, and 2003, with the audio file durations ranging from 10 min to 1 h. The development corpus contains 8 h of data, in 14 audio files recorded from April to July 2003 from the same stations as the training corpus. The evaluation corpus is comprised of 18 audio files recorded from October to December 2004, with a total duration of 10 h. The evaluation corpus contains data from two radio stations ("France Culture" and "Radio Classique") not present in the training or development corpora. There is also a 14-month interval between the recording period of the development and test data (July 2003 to October 2004) of these two corpora. Both data sets present a large variability in audio file durations (10 min to 1 h).

B. Performance Measures

The speaker diarization task performance is measured via an optimum one-to-one mapping between the reference speaker IDs and the hypothesis speaker IDs. The primary metric for the task, referred to as the speaker match error, is the fraction of speaker time that is not attributed to the correct speaker, given the optimum speaker mapping. Another measure is the overall speaker diarization error rate (DER) which includes the missed and false alarm speaker times, thus taking speech/nonspeech detection errors into account [1]. All the measurements mentioned above are illustrated in Fig. 2.

In order to more closely analyze the performance of speaker clustering methods, average frame-level cluster purity and cluster coverage are used as defined by [12]. Cluster purity is defined as the ratio between the number of frames by the dominating speaker in a cluster and the total number of frames in the cluster. Cluster coverage is a dual measure, and accounts for the dispersion of a given speakers data across clusters; for a given speaker, it is defined as the percentage of its frames in the cluster which has most of the data of the speaker. Cluster coverage was also expressed as the purity of reference clusters in [21]. In these experiments, cluster purity and cluster coverage errors are reported.

TABLE I
PURITY, COVERAGE AND OVERALL DIARIZATION ERROR RATES FROM THE **c-std**, **c-bic** AND **c-sid** SYSTEMS ON THE RT-04F AND THE ESTER DEVELOPMENT DATASETS

system	*purity error (%)*	*coverage error (%)*	*overall DER*
RT-04F *dev1* data set			
c-seg	1.0	73.2	N/A
c-std ($\alpha = \beta = 160$)	5.0	28.4	32.3
c-std ($\alpha = \beta = 230$)	9.4	17.9	24.8
c-bic ($\lambda = 5.5$)	2.9	9.8	13.2
c-sid ($\lambda = 3.5, \delta = 0.1$)	2.1	4.2	7.1
RT-04F *dev2* data set			
c-sid ($\lambda = 3.5, \delta = 0.1$)	1.7	3.5	7.6
ESTER development data set			
c-bic ($\lambda = 5.5$)	7.2	10.6	15.8
c-sid ($\lambda = 3.5, \delta = 1.5$)	4.7	5.2	8.0

Cluster purity and coverage measures are complementary, and the speaker match error time can be interpreted as a combination of both. Moreover, it is interesting to note that if the hypothesized speaker for a cluster is always taken to be the majority reference speaker in this cluster, then the speaker match error will be exactly the cluster purity error; it is easy to demonstrate that it is also a lower bound for the match error. Thus, starting with an initial segmentation, the cluster purity error will be the lowest possible match error on this segmentation after performing an agglomerative clustering. The same holds for further clustering of the output of a previous clustering stage, as long as the segment boundaries are not modified.

C. Results on the RT-04F Development Data

The performance at different stages of the system was compared for different system configurations, as presented in Table I. On RT-04F *dev1* data, the initial segmentation **c-seg**, with the minimal duration constraint of 2.5 s per segment, has a purity error of 1% which is also the lowest possible speaker match error. The coverage error of this initial segmentation is of course very high at 73.2%; which means that on average, only about one quarter of the speech for each speaker is located in a single segment.

As expected, the standard partitioner **c-std** in its default configuration provides good purity, but relatively poor coverage, resulting in a high overall diarization error of over 30% on *dev1* data. Setting the penalty α and β to optimize these values reduces this error below 25%. The **c-bic** system also provides a high purity, with a much better coverage (2.9% purity error and 9.8% coverage error), reducing the overall error rate by almost 50%. The **c-sid** system achieves a large decrease of the coverage error with a further small improvement of the purity, resulting in an overall DER of 7.1%, a reduction of almost 50% compared to the **c-bic** system.

A global BIC merging and stop criterion was also tested, but always performed worse than the local BIC criterion in our experiments, as can be seen for **c-bic** system on RT-04F *dev1* in Table II. A similar result was observed in [15]. This result needs further investigation, but may be due to a mismatch between the BIC model and the real distribution of the data. Thus, only the local criterion was used in the remaining experiments.

TABLE II
OVERALL DIARIZATION ERROR FOR **c-bic** SYSTEM ON THE RT-04F *dev1* DATA, AS A FUNCTION OF THE PENALTY WEIGHT λ FOR THE LOCAL AND GLOBAL BIC CRITERION

BIC criterion	λ	*overall DER (%)*	*BIC criterion*	λ	*overall DER (%)*
local	5.0	13.3	global	5.0	16.4
	6.0	12.8		6.0	15.5
	7.0	13.8		7.0	18.2

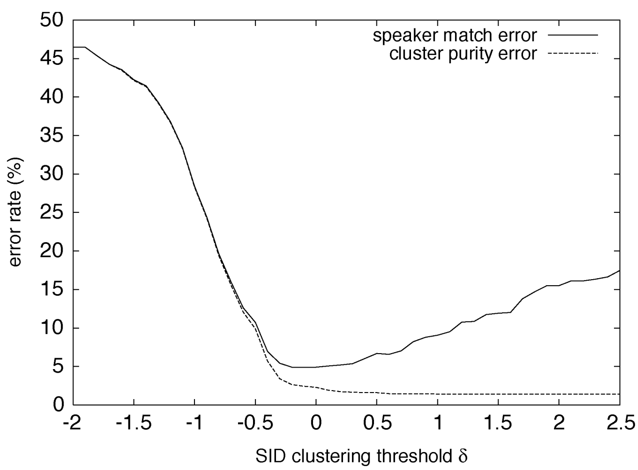

Fig. 3. Speaker match error and purity error rates on RT-04F *dev1* and *dev2* for the **c-sid** system as a function of the SID clustering threshold δ.

The effect of the SID detection threshold δ on the speaker match error and the cluster purity error was measured on both the *dev1* and *dev2* data. A lower threshold reduces the number of final speaker clusters. As shown in Fig. 3, reducing the threshold in the positive range results in a decrease of the match error rate with almost no degradation of the purity. Moreover, there is a large range of thresholds around zero with a low speaker match error. The cluster purity error shows the speaker match error that could be achieved with the best clustering decision, as explained in Section IV-B.

Looking at the performance of the **c-sid** system in more detail, a large variation in the speaker error is observed across shows, as shown in Table III. The speaker error ranges from a low of 0.1% to over 12%. Having only very few speakers (3), the CSPAN show has the lowest speaker error. The ABC and NBC shows have more speakers, occurring in different background conditions, which is more challenging for the diarization system.

D. Results on the ESTER Training and Development Data

For the French ESTER data, SAD was performed using the same American English speech/nonspeech acoustic models as were used for RT-04F plus an additional speech over music model trained on French broadcast news data for a better recognition of the jingles found in the French radio data. The optimal threshold for the SID clustering on development data was $\delta = 1.5$. As can be seen in Table I, the **c-sid** system also has low purity and coverage error rates (4.7% and 5.2%, respectively) on the ESTER development data. A 50% reduction of the overall error rate is gained by adding the **c-sid** system to the **c-bic** system. The δ threshold found for the RT-04F data did not carry over to the ESTER data, and this may be due to the larger variability in show sources, durations, and types observed in ESTER.

The training data was divided into four subsets according to the show duration (10, 15, 20 min, 1 h). As shown in Table IV, different optimal values of SID clustering threshold δ were obtained for each subset. Larger SID clustering thresholds are better for the longer shows. The optimal SID clustering threshold δ on all of the training data is the same as the one for the development data. The show duration was taken here as a rough indicator of the speaker count; however, a more appropriate model of the show type would enable a finer analysis.

E. Results on the Evaluation Data

The trends observed on the development data were confirmed on the RT-04F evaluation data. The results given in Table V show a slight increase in overall diarization error to 17% for the **c-bic** system and to 9.1% for the **c-sid** system. The final SAD post-processing stage gives an improvement of 0.6%, mainly by

TABLE III
PERFORMANCE OF **c-sid** SYSTEM ON THE RT-04F DEVELOPMENT DATASET. SCORES ARE GIVEN FOR MISS (MS), FALSE ALARM (FA), SPEAKER ERROR (SPK), AND OVERALL DIARIZATION ERROR RATE (DER). #REF AND #SYS ARE, RESPECTIVELY, THE REFERENCE AND SYSTEM SPEAKER NUMBER

show	*#REF*	*#SYS*	*MS*	*FA*	*SPK*	*DER*
dev1	121	161	**0.4**	**1.3**	**5.4**	**7.1**
ABC	27	37	1.6	1.3	12.4	15.2
VOA	20	22	0.3	1.2	2.2	3.7
PRI	27	30	0.1	0.9	2.8	3.8
NBC	21	35	0.1	1.1	12.0	13.2
CNN	16	21	0.5	1.4	5.6	7.6
MNB	10	16	0.2	1.8	0.8	2.8
dev2	90	130	**0.5**	**3.1**	**4.1**	**7.6**
CSPAN	3	4	0.3	2.9	0.1	3.3
CNN	17	22	0.6	4.2	5.0	9.8
PBS	27	29	0.1	2.8	7.4	10.3
ABC	23	29	2.1	6.7	12.5	21.2
CNNHL	9	26	0.0	1.4	0.5	1.9
CNBC	11	20	0.2	1.0	0.9	2.1

TABLE IV
RESULTS OF THE DIFFERENT OPTIMAL SID CLUSTERING THRESHOLD δ FOR THE ESTER TRAINING SUBSETS WITH THE DIFFERENT SHOW DURATIONS

training subset	δ	*purity error (%)*	*coverage error (%)*	*speaker match error (%)*
10 min	0.9	1.2	1.5	2.6
15 min	0.9	3.6	0.5	3.8
20 min	1.1	2.2	1.6	3.5
1 hour	1.5	3.5	5.2	7.6
all	**1.5**	**3.0**	**4.7**	**6.7**

TABLE V
PERFORMANCES OF **c-bic**, **c-sid**, AND **p-asr** SYSTEMS ON THE EVALUATION DATA OF RT-04F AND ESTER

system	*missed speech*	*false alarm speech*	*speaker error*	*overall DER*
RT-04F test dataset				
c-bic	0.4	1.8	14.8	17.0
c-sid($\delta = 0.1$)	0.4	1.8	6.9	9.1
p-asr	0.6	1.1	6.8	8.5
ESTER test dataset				
c-bic	0.7	1.0	12.1	13.8
c-sid($\delta = 1.5$)	0.7	1.0	9.8	11.5
post-evaluation result on ESTER test dataset				
c-sid($\delta = 2.0$)	0.7	1.0	7.4	9.1

reducing false alarms in speech detection. As mentioned in [31], the **p-asr** system had the best performance of all the participants of the RT-04F evaluation by a significant margin.

Results on the ESTER evaluation data are given in Table V, with the setting optimized on the development data. The overall diarization error was reduced from 13.8% for the **c-bic** system to 11.5% for the **c-sid** system. The submitted system also had the best performance for this task in the ESTER evaluation [6]. In a post-evaluation experiment, a 20% relative reduction of the overall diarization error was observed for the **c-sid** system with the best *a posteriori* threshold, showing that the error rate is highly dependent on the clustering threshold.

V. Conclusion

In this paper, a multistage architecture for speaker diarization of broadcast news has been described. It builds upon a baseline speaker partitioning system which had been optimized for the automatic transcription task, but the constraints of the speaker diarization task are different. Several modifications to the baseline system have thus been explored. First, the iterative GMM clustering was replaced by an agglomerative BIC clustering, using single full-covariance Gaussian models. A local BIC merging and stop criterion was shown to outperform the global criterion; a similar result was observed in [15]. A second clustering module was applied to the output of the system, relying on techniques used for speaker identification and verification, i.e., acoustic channel normalization, and MAP adaptation of a reference GMM with a large number of Gaussians.

The multistage system was demonstrated to perform much better than the baseline system for the diarization task. On the RT-04F development data, a relative error reduction of over 70% was achieved when compared to the baseline system. This system obtained the best diarization performance in both the NIST RT-04F and the ESTER evaluations by a significant margin. An overall diarization error rate under 10% was obtained on the RT-04F evaluation data, while the error rate of the BIC-based systems were over 15%. Focusing on the speaker error only, the multistage system provides an error reduction of up to 50% relative to a standard BIC clustering system. This dramatic improvement over the baseline system results from several changes: the combination of two different clustering stages, each one focusing on a different acoustic aspect with more complex modeling in the second stage, and the use of acoustic channel normalization methods suited to speaker identification. A system following this architecture recently developed at Cambridge University demonstrated similar improvements [32], where it was observed that a very important part of the gain was obtained by the feature warping normalization.

Several remaining issues need further investigation in order to improve the robustness and the efficiency of the system. It was observed that the clustering threshold needs to be set according to the type of the audio document, and that the system still has a large variability across individual shows. Only with a large amount of files can statistically consistent results be obtained. This is especially important since the speaker error does not provide a stable and continuous measure of a clustering system: splitting a speaker in two classes, which is a single decision, results in doubling of the error rate for this speaker.

Finally, most speaker diarization systems rely on a purely acoustic segmentation and clustering, whereas an essential part of the information in speech is of a linguistic nature, and obviously in TV and radio shows, most speakers are presented and identified. Combining the acoustic information with the linguistic layer as explored in [33] would improve the robustness of a speaker diarization system and make it more exploitable by a human reader.

References

[1] NIST. (2004, Aug.) Fall 2004 Rich Transcription (RT-04F) evaluation plan, Gaithersburg, MD. [Online] Available: http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf

[2] S. E. Tranter, K. Yu, D. A. Reynolds, G. Evermann, D. Y. Kim, and P. C. Woodland, “An investigation into the interactions between speaker diarization systems and automatic speech transcription,” Eng. Dept., Cambridge Univ., Cambridge, U.K., Tech. Rep. CUED/F-INFENG/TR-464, Oct. 2003.

[3] NIST. (2003, Feb.) The Rich Transcription Spring 2003 (RT-03S) evaluation plan, Gaithersburg, MD. [Online] Available: http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf

[4] ——, (2004, Feb.) Spring 2004 (RT-04S) Rich Transcription meeting recognition evaluation plan, Gaithersburg, MD. [Online] Available: http://www.nist.gov/speech/tests/rt/rt2004/spring/documents/rt04s-meeting-eval-plan-v1.pdf

[5] G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri, “The ESTER evaluation campaign of rich transcription of French broadcast news,” in *Proc. Lang. Evaluation Resources Conf. (LREC 2004)*, Lisbon, Portugal, May 2004, pp. 885–888.

[6] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, “The ESTER phase II evaluation campaign for the rich transcription of French broadcast news,” in *Proc. 9th Eur. Conf. Speech Communication and Technology (ISCA Interspeech)*, Lisbon, Portugal, Sep. 2005, pp. 1149–1152.

[7] Y. Moh, P. Nguyen, and J.-C. Junqua, “Toward domain independent speaker clustering,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, China, Apr. 2003, pp. II-85–II-88.

[8] M. Ben, M. Betser, F. Bimbot, and G. Gravier, “Speaker diarization using bottom-up clustering based on a parameter-derived distance between GMMs,” in *Proc. Int. Conf. Spoken Language Process.*, Jeju, Korea, Oct. 2004, pp. 1125–1128.

[9] M. Siegler, U. Jain, B. Raj, and R. Stern, “Automatic segmentation and clustering of broadcast news audio,” in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, Feb. 1997, pp. 97–99.

[10] S. Chen and P. Gopalakrishnan, “Speaker, environment, and channel change detection and clustering via the Bayesian information criterion,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, Feb. 1998, pp. 127–132.

[11] T. Hain and P. C. Woodland, “Segmentation and classification of broadcast news audio,” in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, Nov. 1998, pp. 2727–2730.

[12] J.-L. Gauvain, L. Lamel, and G. Adda, “Partitioning and transcription of broadcast news data,” in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, Dec. 1998, pp. 1335–1338.

[13] S. Meignier, J.-F. Bonastre, and S. Igounet, “E-HMM approach for learning and adapting sound models for speaker indexing,” in *2001: A Speaker Odyssey. Proc. Speaker Recognition Workshop (ISCA, Odyssey 2001)*, Chania, Crete, Jun. 2001, pp. 175–180.

[14] J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *Proc. Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands, Nov. 2003, pp. 411–416.

[15] S. E. Tranter and D. A. Reynolds, “Speaker diarization for broadcast news,” in *Proc. 2004: A Speaker Odyssey. The Speaker Recognition Workshop*, Toledo, Spain, May 2004, pp. 337–344.

[16] J.-L. Gauvain and C. H. Lee, “Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[17] J.-L. Gauvain, L. Lamel, and G. Adda, "Audio partitioning and transcription for broadcast data indexation," *Multimedia Tools Applicat.*, vol. 14, pp. 187–200, 2001.

[18] ——, "The LIMSI broadcast news transcription system," *Speech Commun.*, vol. 37, no. 1–2, pp. 89–108, 2002.

[19] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[20] P. Delacourt and C. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Commun.*, vol. 32, pp. 111–126, 2000.

[21] M. Cettolo, "Segmentation, classification, and clustering of an Italian broadcast news corpus," in *Proc. Content-Based Multimedia Inf. Access Conf.*, Paris, France, Apr. 2000, pp. 372–381.

[22] M. Cettolo, M. Vescovi, and R. Rizzi, "Evaluation of BIC-based algorithms for audio segmentation," *Comput. Speech Lang.*, vol. 19, pp. 147–170, 2005.

[23] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, Nov. 2000, pp. 1423–1426.

[24] J. Schroeder and J. Campbell, Eds., *Digital Signal Processing (DSP), a Review Journal—Special Issue on NIST 1999 Speaker Recognition Workshop*. Norwell, MA: Academic, 2000.

[25] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, China, 2003, pp. II-49–II-52.

[26] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. 2001: A Speaker Odyssey. The Speaker Recognition Workshop*, Chania, Crete, Jun. 2001, pp. 213–218.

[27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing (DSP), a Review Journal—Special Issue on NIST 1999 Speaker Recognition Workshop*, vol. 10, pp. 19–41, 2000.

[28] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. J. McLaughlin, and M. A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proc. Int. Conf. Spoken Language Process.*, Sydney, Australia, Nov. 1998, pp. 3193–3196.

[29] L. Nguyen, S. Abdou, M. Afify, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang, L. Lamel, J.-L. Gauvain, G. Adda, H. Schwenk, and F. Lefevre, "The 2004 BBN/LIMSI 10xRT English broadcast news transcription system," in *Proc. DARPA RT04'S*, Palisades, NY, Nov. 2004, pp. 33-1–33-7.

[30] X. Zhu, C. Barras, S. Meignier, and J-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 2441–2444.

[31] D. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. Int. Conf. Acoust., Speech, Signal Process*, Philadelphia, PA, Mar. 2005, pp. 953–956.

[32] R. Sinha, S. Tranter, M. Gales, and P. Woodland, "The Cambridge University March 2005 speaker diarization system," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 2437–2440.

[33] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcripts," in *Proc. Int. Conf. Spoken Language Process.*, Jeju Island, Korea, Oct. 2004, pp. 1272–1275.

Claude Barras received the engineering degree from the Ecole Supérieure d'Electricité Supélec, Gif-sur-Yvette, France, and the Ph.D. degree in computer science from the University Pierre-et-Marie-Curie, Paris, France, in 1996.

Since 2001, he has been an Associate Professor at the University Paris-Sud, France, and works in the Spoken Language Processing Group, LIMSI-CNRS, Orsay, France. His research fields include speech recognition, speech annotation, and speaker recognition.

Xuan Zhu received the B.Sc. degree in computer science from Central China Normal University, Wuhan, China, in 2000 and the M.Sc. degree in computer science from the University Paris-Sud, Paris, France, in 2003. She is currently pursuing the Ph.D. degree in the Spoken Language Processing Group at LIMSI-CNRS, Orsay, France, and works on the topic of speaker diarization under the supervision of Dr. J.-L. Gauvain and Dr. C. Barras.

Sylvain Meignier received the Ph.D. degree in computer science from the Université d'Avignon et des Pays de Vaucluse, Avignon, France, in 2002. His work was about speaker indexing of audio documents.

In 2003, he was with LIMSI-CNRS, Orsay, France, in the Spoken Language Processing Group as a Researcher. Since 2004, he has been an Associate Professor at the Université du Maine, Le Mans, France, where he works in the LIUM laboratory on speech processing.

Jean-Luc Gauvain (M'98) received the Ph.D. degree in electronics from the University of Paris XI, Paris, France, in 1982.

He is a Senior Researcher at the CNRS where he is head of the LIMSI Spoken Language Processing Group (http://www.limsi.fr/~gauvain). His primary research centers on large-vocabulary continuous speech recognition and audio indexing. His research interests also include conversational interfaces, speaker identification, language identification, and speech translation. He has participated in many speech-related projects both at the French National and European levels and has led the LIMSI participation in the DARPA/NIST organized evaluations since 1992, most recently for the transcription of broadcast news data and of conversational speech. He has over 220 publications.

Dr. Gauvain received the 1996 IEE SPS Best Paper Award in Speech Processing and the 2004 ISCA Best Paper Award for a paper in the *Speech Communication Journal*. He was a member of the IEEE Signal Processing Societys Speech Technical Committee from 1998 to 2001, and is currently Co-Editor-In-Chief of the *Speech Communication Journal*.