Speaker Diarization: from Broadcast News to Lectures

X. Zhu, C. Barras, L. Lamel, and J-L. Gauvain*

LIMSI-CNRS, BP 133
91403 Orsay Cedex, France

Abstract. This paper presents the LIMSI speaker diarization system for lecture data, in the framework of the Rich Transcription 2006 Spring (RT-06S) meeting recognition evaluation. This system builds upon the baseline diarization system designed for broadcast news data. The baseline system combines agglomerative clustering based on Bayesian information criterion with a second clustering using state-of-the-art speaker identification techniques. In the RT-04F evaluation, the baseline system provided an overall diarization error of 8.5% on broadcast news data. However since it has a high missed speech error rate on lecture data, a different speech activity detection approach based on the log-likelihood ratio between the speech and non-speech models trained on the seminar data was explored. The new speaker diarization system integrating this module provides an overall diarization error of 20.2% on the RT-06S Multiple Distant Microphone (MDM) data.

1 Introduction

Audio diarization is the process of partitioning an input audio stream into homogeneous segments according to their specific audio source such as speaker identity, category of music, background noise or channel conditions. Speech activity detection (SAD) is the simplest case of diarization, which just divides the audio data into speech/non-speech segments. Speaker diarization, also referred to as speaker segmentation and clustering, is a more complicated task than audio diarization, and needs to determine segments consisting of the speech from only one speaker and associate speech segments from the same speaker. SAD is a very useful preprocessing step for many audio technologies such as automatic speech recognition, speaker identification and verification, speaker localization etc. Speaker diarization has been used in Automatic Speech Recognition (ASR) to carry out unsupervised speaker adaptation, where the amount of data available for the adaptation can be increased by clustering segments from the same speaker. Speaker diarization can also improve the readability of an automatic transcription by structuring the audio stream into speaker turns and is of interest for the indexing of multimedia documents.

The challenges for the speaker diarization task derive from the varied data types: Broadcast News (BN), telephone conversations and meeting recordings. Most research efforts on speaker diarization have focused on the broadcast news domain [1, 2]. Recently there has been strong interest in the meeting domain [3, 4], which poses more

* This work was partially financed by the European Commission under the FP6 Integrated Project IP 506909 CHIL

difficulties for the speaker diarization task. The speech in the meeting is completely spontaneous, with frequent periods of overlapping speech and a large number of silence segments for any given speaker. Meetings are usually recorded using different types of microphones located at various positions in the room, providing multiple audio files with different signal qualities for the same meeting. The use of the distant microphones also makes the audio signal more noisy than many of the broadcast news recordings.

In the Rich Transcription 2006 Spring (RT-06S) meeting recognition evaluation [5], the task was divided into two sub-domains: conference room meetings and lecture room meetings (seminar-like meetings). Compared with the conference data, the lecture meetings have less interaction between the participants, and typically consist of a presentation from a lecturer followed by a question/answering session or discussion period.

LIMSI participated in the speech activity detection and speaker diarization tasks of the RT-06S evaluation, focusing on the lecture data. The LIMSI multi-stage speaker diarization system developed for BN data [6] was adapted to the lecture data, especially the SAD module. This modified system was tested on far-field conditions: the Multiple Distant Microphone (MDM), Single distant Microphone (SDM) and Multiple Mark III Microphone Array (MM3A). As defined by this evaluation, no a priori knowledge of the speaker's voice or even the number of speakers is available for the distant microphone conditions, and thus only a relative and recording-internal speaker identification is produced by the system.

The remainder of this paper is organized as follows: Section 2 describes the baseline speaker diarization system for broadcast news data, and Section 3 presents the log-likelihood based speech activity detection adapted to the lecture data. The experimental results are presented in Section 4, followed by some conclusions.

2 Baseline BN diarization system

The baseline speaker diarization system developed for Broadcast News combines an agglomerative clustering based on Bayesian information criterion (BIC) with a second clustering stage which uses state-of-the-art speaker identification (SID) methods. It obtains good performance on BN data, achieving an overall speaker diarization error of 8.5% on RT-04F evaluation data [7]. The primary system is structured as follows:

2.1 Feature extraction

Mel frequency cepstral parameters are extracted from the speech signal every 10 ms using a 30 ms window on a 0-8kHz band. The 38 dimensional feature vector consists of 12 cepstral coefficients, Δ and Δ-Δ coefficients plus the Δ and Δ-Δ log-energy. Acoustic vector normalization is only performed in the SID clustering stage.

2.2 Speech Activity Detection (SAD)

Speech is extracted from the signal with a Viterbi decoding using Gaussian Mixture Models (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 hour of the specific type of data, selected from English Broadcast News data distributed by the Linguistic Data Consortium (LDC).

2.3 Initial segmentation

The segmentation process consists of finding segment boundaries that correspond to the instantaneous speaker change points. The initial 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 of 5 seconds, similar to the KL2 metric based segmentation [8]. Each window is modeled by a single diagonal Gaussian using the static features (i.e., only the 12 cepstral coefficients plus the energy). More precisely, the Gaussian divergence measure is defined as:

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

with $s_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ and Σ_i diagonal, $i \in \{1, 2\}$. The detection threshold was optimized on the training data in order to provide acoustically homogeneous segments.

2.4 Viterbi resegmentation

An 8-component GMM with a diagonal covariance matrix is trained on each segment resulting from the initial segmentation, the boundaries of the speech segments detected by the SAD module are then refined using a Viterbi segmentation with this set of GMMs.

2.5 BIC clustering

An initial cluster c_i is modeled by a single Gaussian with a full covariance matrix Σ_i estimated on the n_i acoustic frames of each segment output by the Viterbi resegmentation. The BIC criterion [9] is used both for the inter-cluster distance measure and the stop criterion. It is defined as:

$$\Delta BIC = (n_i + n_j) \log |\Sigma| - n_i \log |\Sigma_i| - n_j \log |\Sigma_j| - \lambda \frac{1}{2} (d + \frac{1}{2} d(d+1)) \log n \quad (2)$$

where d is the dimension of the feature vector space, $n = n_i + n_j$ and λ weights the BIC penalty. At each step the two nearest clusters are merged, and the ΔBIC values between this new cluster and the remaining clusters are computed. This clustering procedure stops when all ΔBIC are greater than zero.

2.6 SID clustering

After the BIC clustering stage, speaker recognition methods [11, 12] are used to improve the quality of the speaker clustering. Feature warping normalization [13] is performed on each segment using a sliding window of 3 seconds in order to map the cepstral feature distribution to a normal distribution and reduce the non-stationary effects of the acoustic environment. The GMM of each remaining cluster is obtained by maximum a posteriori (MAP) adaptation [15] of the means of the matching Universal Background Model (UBM [14]). For each gender and channel condition (wide band, narrow band) combination, an UBM with 128 diagonal Gaussians is trained on the corresponding subset of 1996/1997 English Broadcast News data. Then a second stage of agglomerative clustering is performed using the cross log-likelihood ratio as in [16]:

$$\mathcal{S}(c_i, c_j) = \frac{1}{n_i} log \frac{f(x_i|M_j)}{f(x_i|B)} + \frac{1}{n_j} log \frac{f(x_j|M_i)}{f(x_j|B)} \quad (3)$$

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. The clustering stops when the cross log-likelihood ratio between all clusters is below a given threshold δ optimized on the development data (see Section 4.1).

2.7 SAD post-filtering

The word segmentation output by the LIMSI Broadcast News Speech-To-Text system [17] is used in a post-processing stage to filter out short-duration silence segments that were not removed by the initial speech activity detection step. Only inter-word silences longer than 1 second are filtered out, this value having been determined empirically.

3 Speech activity detection for lectures

The LIMSI RT-06S speaker diarization system for lecture data was built upon the broadcast news diarization system. Initial results on the development data with the baseline system had a high speech activity detection error, especially with a lot of missed speech, therefore a different approach for SAD was explored. One weakness of the standard Viterbi decoding is the lack of temporal control for each model. A transition penalty can be used to control the size of the segments, but as the level of noise increases, the likelihood of the speech model will decrease and thus the shortest speech segments will be discarded. Instead of setting a minimal likelihood level for switching from one model to the other, it is easier to choose a minimal duration for speech and non-speech segments.

We designed a simple speech activity detector based on the log-likelihood ratio (LLR) between the speech and non-speech models, and replaced the Viterbi decoding with a simple smoothing of the LLR followed by a decision module. More precisely:

– for each frame x_i, the LLR r_i between the speech and non-speech models λ_S and λ_N is computed taking into account their prior probabilities $P(S)$ and $P(N)$:

$$r_i = \log f(x_i|\lambda_S)P(S) - \log f(x_i|\lambda_N)P(N)$$

– two adjacent smoothing windows with a duration of $w = 100$ frames (i.e. 1 second) sliding over the signal are used for the detection of speech and non-speech transitions. A transition is possible when the sign of the averaged LLR in the left and right windows changes around the current frame:

$$s_i^+ . s_i^- < 0 \quad \text{with} \quad s_i^+ = \frac{1}{w} \sum_{j=i+1}^{i+w} r_j \quad \text{and} \quad s_i^- = \frac{1}{w} \sum_{j=i-w}^{i-1} r_j$$

– for a set I consisting of contiguous candidate transitions, the position of the transition is chosen at the maximum of difference between the averaged ratio of the left and right windows:

$$i^* = \underset{i \in I}{\operatorname{argmax}} |s_i^+ - s_i^-|$$

The GMMs for speech and non-speech were trained on about 2 hours of far-field data from seminars recorded at the University of Karlsruhe (UKA).

4 Experimental results

In RT-06S lecture room evaluation, results were submitted for the SAD and speaker diarization tasks on three audio input conditions: MDM, SDM and MM3A. The configurations of BIC clustering and SID clustering were optimized on the development data. All experiments were carried out with the BIC penalty weight $\lambda = 3.5$ and the SID threshold $\delta = 0.5$.

4.1 Performance measures and databases descriptions

The speaker diarization task performance is measured via an optimum mapping between the reference speaker IDs and the hypotheses. This is the same metric as was used to evaluate the performance on BN data. The primary metric for the task, referred to as the speaker error, is the fraction of speaker time that is not attributed to the correct speaker, given the optimum speaker mapping. In addition to this speaker error, the overall speaker diarization error rate (DER) also includes the missed and false alarm speaker times. The SAD task performance is evaluated by summing the missed and false alarm speaker error. In RT-06S evaluation, the metrics are calculated over all the speech, including the overlapping speech. The DER restricted to non-overlapping speech segments is also given for comparison purposes.

The experiments were conducted on the NIST RT-06S evaluation data comprised of lectures provided by the CHIL (Computer in the Human Interaction Loop) consortium. The data were collected at 5 of the CHIL partner sites: AIT (Athens Information Technology), IBM, ITC (Istituto Trentino di cultura), UKA and UPC (Universitat Politecnica de Catalunya). The development dataset (dev) consists of all seminars used as

RT-05s evaluation data, plus an additional seminar from UKA and four seminars from AIT, IBM, ITC and UPC one each. The evaluation dataset (eval) is composed of 38 seminar segments each lasting about 5 minutes.

4.2 Audio input selection

For the MDM evaluation condition, a single microphone signal randomly chosen from the available MDM channels and different from the channel selected for the SDM condition was used as the input to the speaker diarization system. Because the same microphone type is used for the MDM and SDM conditions, no individual development was carried on SDM condition, i.e. the same configuration for the speaker diarization system is adopted for both conditions. The microphone channels used for the MDM and SDM conditions are detailed in Table 1. For MM3A evaluation condition, the beamformed multiple mark III microphone array data provided by UKA was used as the input of the speaker diarization system.

Table 1. Channel selection for the MDM and the SDM conditions for the dev and eval data.

Dataset	*Condition*	*AIT*	*IBM*	*ITC*	*UKA*	*UPC*
dev	MDM	mic05	Audio_17	Table-1	TableTop-1	channel15
eval	MDM	mic06	Audio_17	Table-2	TableTop-1	channel16
eval	SDM	mic05	Audio_19	Table-1	Table-2	channel15

4.3 RT-06S MDM development results

The performances of the speaker diarization systems integrating different SAD modules are summarized in Table 2. The "vit-bn" system uses Viterbi decoding with 5 GMMs (64 Gaussians) for speech, noisy speech, speech over music, pure music, and silence, each trained on one hour of BN data. This baseline speaker diarization system is the same system as was used in RT-04F evaluation for BN data. The "vit-bn+mt" system uses Viterbi decoding with GMMs trained on the BN data plus 2 GMMs (256 Gaussians) for speech and non-speech trained on 2 hours of far-field data from the UKA seminars. The "vit-mt" system uses Viterbi decoding only with speech and non-speech models trained on lecture data. The "slr-mt" system uses the smoothed LLR-based SAD method with a prior probability of 0.2 for the non-speech model and 0.8 for the speech model. As can be seen in Table 2, Viterbi SAD using the models trained on both BN and lecture data have very high missed speech error rates (ranging from 18% to 14%) on the MDM development data. The log-likelihood based SAD substantially reduces this error (2.7% missed speech error) with limited increase in false alarm speech error. Compared with the baseline speaker diarization system, a relative DER reduction of 33% is obtained by the system using the smoothed LLR-based SAD.

Table 3 gives the speaker diarization results on the MDM development data when the number of Gaussians for the speech and the non-speech models used in the smoothed

Table 2. Speaker diarization errors on the MDM development data for different SAD modules.

System	*Missed speech (%)*	*False alarm speech (%)*	*Speaker error (%)*	*Overlap DER (%)*
vit-bn (baseline)	18.2	3.0	9.0	30.2
vit-bn+mt	19.3	2.9	8.7	31.0
vit-mt	14.2	3.7	12.4	30.2
slr-mt	**2.7**	**6.1**	**11.7**	**20.5**

LLR-based SAD are varied. These results are obtained with a prior probability of 0.4 for the non-speech model and 0.6 for the speech model. There are no gains of the overall diarization error when the number of Gaussians is increased from 256 to 512 on the MDM development data.

Table 3. Results varying the number of Gaussians for the speech and non-speech models on the MDM development data.

nb. Gaussians	*Missed speech (%)*	*False alarm speech (%)*	*Speaker error (%)*	*Overlap DER (%)*
64	9.5	4.0	11.0	24
128	9.5	3.7	11.0	24
256	**7.8**	**4.2**	**11.0**	**23**
512	7.7	4.2	11.1	23

The effect of the prior probabilities for speech and non-speech used in LLR-based SAD was also studied. The results presented in Table 4 are obtained with 256-component GMMs used for each model. Because it is important for automatic speech transcription to reject the least amount of speech as possible, a higher prior probability for the speech model is preferred relative to the non-speech model. As shown in Table 4, using a prior probability of 0.2 for the non-speech model and 0.8 for the speech model provides the best results for both speech activity detection (8.8% SAD error) and speaker diarization (20.5% DER).

Table 4. Results obtained by using different prior probabilities for the speech and non-speech models on the MDM development data.

P(N):P(S)	*Missed speech (%)*	*False alarm speech (%)*	*Speaker error (%)*	*Overlap DER (%)*
0.1:0.9	1.0	9.5	12.0	22.4
0.2:0.8	**2.7**	**6.1**	**11.7**	**20.5**
0.3:0.7	5.2	5.0	11.3	21.5
0.4:0.6	7.8	4.2	11.0	23.0

After the experiments on the MDM development data, the configuration of the log-likelihood based SAD system is optimized as: a prior probability of 0.2 for the non-speech model and 0.8 for the speech model with 256-component GMMs used for both models. The performance of the speaker diarization system using the LLR-based SAD module is presented in Table 5, where the result is given for the individual seminar having the corresponding reference released by NIST. As shown in Table 5, the average DER of 20.5% masks the large variation across seminars. Normally lower overall diarization error can be obtained on the seminars with only one speaker, but for "UKA_20041124_A_Segment2" seminar, a very high false alarm speech error of about 150% is produced by the LLR-based SAD module. After listening to the audio file, we found that many speech segments are missing in the reference transcription, this may be because the speech signal was not recorded on the microphone channel chosen for the manual reference transcription.

In order to analyze the variation in system performance, we calculated the ratio between the speech time from the main speaker (who spoke the most in the seminar) and the total seminar duration on all the seminars in Table 5 except the "UKA_20041124_A-_Segment2" seminar. Figure 4.3 shows that the speaker diarization system provides lower overall diarization error on seminars where the main speaker spoke for more than 80% of the seminar duration. Moreover a correlation between the DER and the dominant speaker duration ratio is apparent clearly; consistent with the observations reported in [18].

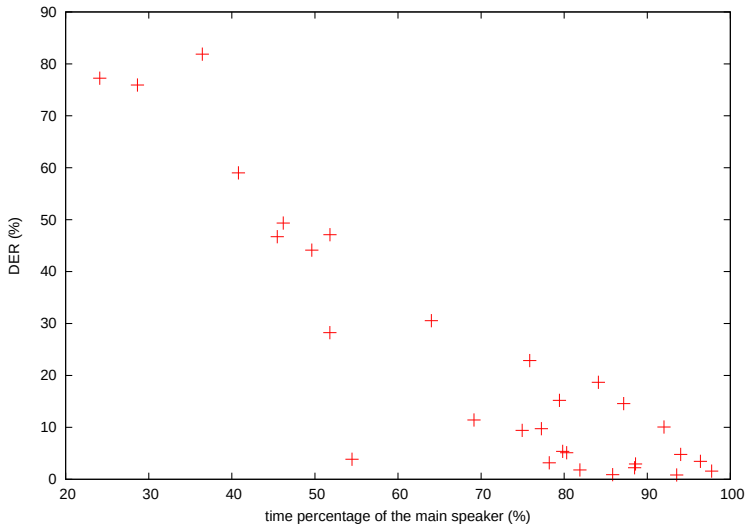

Fig. 1. Overall speaker diarization error on the MDM development data as a function of the time percentage of speech for the main speaker during the seminar.

Table 5. Results by seminar in the MDM development dataset, "REF" represents the number of speakers in the reference transcriptions.

Seminar	*Missed speech (%)*	*False alarm speech (%)*	*Speaker error (%)*	*Overlap DER (%)*	REF
AIT_20050726_Segment1	0.9	11.3	10.7	22.9	4
IBM_20050824_Segment1	2.6	0.9	1.6	5.1	2
ITC_20050429_Segment1	2.3	4.1	8.1	14.6	3
UKA_20041123_A_Segment1	0.0	0.9	0.0	0.9	1
UKA_20041123_A_Segment2	0.9	0.0	29.6	30.6	2
UKA_20041123_B_Segment2	7.2	35.0	4.6	46.7	3
UKA_20041123_C_Segment1	0.6	1.6	0.0	2.2	1
UKA_20041123_C_Segment2	1.7	5.7	4.1	11.4	3
UKA_20041123_D_Segment1	7.9	1.3	0.8	10.1	1
UKA_20041123_D_Segment2	1.6	75.5	4.8	81.9	2
UKA_20041123_E_Segment1	1.4	0.8	7.6	9.8	2
UKA_20041123_E_Segment2	3.5	5.1	6.6	15.2	2
UKA_20041124_A_Segment1	1.7	7.6	0.1	9.4	1
UKA_20041124_A_Segment2	3.2	149.9	4.5	157.6	1
UKA_20041124_B_Segment1	0.2	1.4	0.3	1.8	1
UKA_20041124_B_Segment2	1.9	3.2	44.2	49.3	4
UKA_20050112_Segment1	4.5	0.3	0.0	4.8	1
UKA_20050112_Segment2	10.2	1.2	7.2	18.7	3
UKA_20050126_Segment1	0.3	2.9	0.0	3.2	1
UKA_20050127_Segment1	1.3	0.2	0.1	1.6	1
UKA_20050128_Segment1	2.3	1.1	0.0	3.5	1
UKA_20050128_Segment2	3.0	1.7	39.5	44.1	5
UKA_20050202_Segment2	8.8	11.0	57.4	77.2	7
UKA_20050209_Segment1	2.5	1.4	0.0	3.9	1
UKA_20050209_Segment2	11.4	11.8	52.7	75.9	4
UKA_20050310_A_Segment1	0.5	1.7	0.8	3.0	1
UKA_20050310_A_Segment2	1.0	4.1	53.9	59.0	4
UKA_20050310_B_Segment1	0.2	0.6	0.0	0.8	1
UKA_20050314_Segment1	3.1	1.8	0.5	5.4	1
UKA_20050314_Segment2	6.0	3.3	19.0	28.3	4
UPC_20050601_Segment1	2.7	24.4	20.0	47.1	3
all	2.7	6.1	11.7	20.5	-

4.4 RT-06S evaluation results

The RT-06S evaluation results are given in Table 6. For the MDM and SDM conditions, system tuning used the same development data, and therefore identical configurations are used for both conditions. The system performance is quite similar to that obtained on the MDM development data with an overlap overall diarization error of 21.5%. For the SDM audio input condition, the overlap DER is increased to 24.5%. This increase of the diarization error comes mainly from the SAD error, due to the different quality of the microphone channels used for the MDM and SDM conditions.

For the MM3A contrast condition, the system configuration was optimized on the beamformed development data. Since no adaptation of the SAD acoustic models is performed on the beamformed data, a slightly higher diarization error of 25.9% is obtained for the MM3A condition relative to the MDM condition.

Table 6. Evaluation results for SAD and speaker diarization for the MDM, SDM and MM3A conditions.

Condition	*nb. Gaussians*	*P(S)*	*Overlap SAD error (%)*	*Overlap DER (%)*	*Non-overlap DER (%)*
MDM	256	0.8	9.0	21.5	20.2
SDM	256	0.8	12.4	24.5	23.2
MM3A	128	0.6	11.5	25.9	24.7

5 Conclusions

The work at LIMSI related to speech activity detection and speaker diarization in the framework of the RT-06S meeting recognition evaluation was reported in this paper. Our speaker diarization system for the lecture task builds upon a baseline multi-stage system developed for broadcast news. The main modification is the use of a smoothed log-likelihood ratio based SAD with acoustic models adapted to the lecture data. This SAD was demonstrated to perform much better than the baseline Viterbi SAD. On the MDM development data, the LLR-based SAD provides a significant reduction of the SAD error up to 58% relative to Viterbi SAD, and in particular reduces the missed speech error. Concerning the speaker diarization performance, the diarization system using the LLR-based SAD gives an overall error of 20% , compared to the 30% overall error obtained with the baseline system. On the evaluation data, the RT-06S speaker diarization system provides an overlap overall diarization error of 21.5% on the MDM condition, with a small increase in the overlap DER to 24.5% for the SDM condition and a higher error of 25.9% for the MM3A condition. The robustness of the speaker diarization system depends a lot on the data domain. The combination of BIC clustering and SID clustering is very effective on the BN data and provides 8.5% non-overlapping overall diarization error on RT-04F evaluation data. A relatively higher non-overlapping DER of 20.2% is obtaind on the MDM lecture data. This decrease of the speaker diarization performance may derive from the lower signal quality of the lecture data.

Our future work will focus on the improvement of the robustness of the speaker diarization system. Efficiently using information from all of the available MDM microphone channels is another important research direction.

References

1. S. E. Tranter and D. A. Reynolds, "Speaker diarisation for broadcast news," in *Proc. ISCA Speaker Recognition Workshop Odyssey 2004*, Toledo, Spain, May 2004.
2. C. Barras, X. Zhu, S. Meignier and J-L. Gauvain, "Multi-Stage Speaker Diarization of Broadcast News," to appear in *The IEEE Transactions on Audio, Speech and Language Processing*, September, 2006 (to appear).
3. X. Anguera, C. Wooters, B. Peskin and M. Aguilo, "Robust Speaker Segmentation for Meeting: The ICSI-SRI Spring 2005 Diarization System," in *MLMI 2005 Meeting Recognition Workshop*, Edinburgh, UK, July 2005.
4. D. Istrate, C. Fredouille, S. Meignier, L. Besacier and J-F. Bonastre, "NIST RT05S evaluation: pre-processing techniques and speaker diarization on multiple microphone meetings," in *MLMI 2005 Meeting Recognition Workshop*, Edinburgh, UK, July 2005.
5. NIST, "Spring 2006 Rich Transcription (RT-06S) Meeting Recognition Evaluation Plan," http://www.nist.gov/speech/tests/rt/rt2006/spring/docs/rt06s-meeting-eval-plan-V2.pdf , February, 2006.
6. X. Zhu, C. Barras, S. Meignier, and J-L. Gauvain, "Combining Speaker Identification and BIC for Speaker Diarization," in *ISCA Interspeech'05*, Lisbon, September 2005, pp. 2441–2444.
7. NIST, "Fall 2004 Rich Transcription (RT-04F) evaluation plan," http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf , August 2004.
8. M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic segmentation and clustering of broadcast news audio," in *the DARPA Speech Recognition Workshop*, Chantilly, USA, Feb. 1997.
9. S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, USA, Feb. 1998.
10. M. Cettolo, "Segmentation, classification and clustering of an Italian broadcast news corpus," in *Conf. on Content-Based Multimedia Information Access (RIAO 2000)*, Paris, April 2000.
11. J. Schroeder and J. Campbell, Eds., *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, Academic Press, 2000.
12. C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *IEEE ICASSP 2003*, Hong Kong, 2003.
13. J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Speaker Recognition Workshop Odyssey 2001*, Chania, Crete, June 2001, pp. 213–218.
14. 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, no. 1-3, pp. 19–41, 2000.
15. J.-L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2(2), pp. 291–298, April 1994.
16. 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. of International Conf. on Spoken Language Processing (ICSLP'98)*, 1998.

17. 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 *DARPA RT04'S*, Palisades, NY, Nov 2004.
18. N. Mirghafori and C. Wooters, "Nuts and Flakes: A Study of Data Characteristics in Speaker Diarization," in *IEEE ICASSP 2006*, pp. 1017–1020, Toulouse, May 2006.