Multi-Stage Speaker Diarization for Conference and Lecture Meetings

X. Zhu[1,2], C. Barras[1,2], L. Lamel[1], and J-L. Gauvain[1*]

[1]Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, France
[2]Univ Paris-Sud, F-91405, Orsay, France

Abstract. The LIMSI RT-07S speaker diarization system for the conference and lecture meetings is presented in this paper. This system builds upon the RT-06S diarization system designed for lecture data. The baseline system combines agglomerative clustering based on Bayesian information criterion (BIC) with a second clustering using state-of-the-art speaker identification (SID) techniques. Since the baseline system provides a high speech activity detection (SAD) error around of 10% on lecture data, some different acoustic representations with various normalization techniques are investigated within the framework of log-likelihood ratio (LLR) based speech activity detector. UBMs trained on the different types of acoustic features are also examined in the SID clustering stage. All SAD acoustic models and UBMs are trained with the forced alignment segmentations of the conference data. The diarization system integrating the new SAD models and UBM gives comparable results on both the RT-07S conference and lecture evaluation data for the multiple distant microphone (MDM) condition.

1 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 background acoustic and channel conditions. The aim of the speaker diarization is to provide a set of speech segments, where each segment is bounded by speaker change or speech/non-speech change points and labeled with the identity of the speaker engaging in the corresponding speech.

Speaker diarization was evaluated for Broadcast news data in English up to 2004, and the meeting domain became the main focus of NIST evaluations since 2005. Beside the conference room and lecture room sub-domains, a new type of recordings has been introduced into the NIST 2007 Spring meeting recognition evaluation [1], that is, the recordings of coffee breaks. Since the lecture room and coffee break excerpts were extracted from different parts of the same meetings, both sub-domains have the same sensor configurations but these are different from the conference meeting sub-domain. Although these three types of data have different styles of participant interaction, the

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

RT-07S lecture evaluation data contains more interactions between meeting participants than the previous lecture meeting data.

In the RT-07S evaluation, LIMSI participated to the speaker diarization task on the conference and lecture data and developed a general diarization system for both types of meeting data. As the RT-06S lecture diarization system had a high SAD error that in turn affects strongly final speaker diarization performance, some new acoustic models trained using the forced alignment transcriptions and several different feature normalization techniques are employed to reduce the speech activity detection error. In order to ameliorate the SID clustering performance, several new UBMs were also constructed in a similar manner. For the MDM audio input condition, the RT-07S diarization system uses the beamformed signals generated from the ICSI delay&sum signal enhancement system [2] instead of selecting randomly one channel from all available MDM channels as was done by LIMSI RT-06S lecture diarization system.

The remainder of this paper is organized as follows: Section 2 describes the baseline speaker diarization system for lecture data, and Section 3 presents a proposed acoustic representation for speech activity detection. The experimental results are presented in Section 4.

2 Baseline lecture diarization system

The baseline lecture speaker diarization system combines an agglomerative clustering based on Bayesian information criterion (BIC) with a second clustering stage which uses state-of-the-art speaker identification methods. This combined clustering technique was first proposed for Broadcast News data [3, 4] and was proven to be effective also for lectures data [5]. This baseline system processes 38 dimensional feature vectors consisting of 12 cepstral coefficients, Δ and Δ-Δ coefficients plus the Δ and Δ-Δ log-energy. The primary system modules are:

- Speech activity detection (SAD): speech is extracted from the signal by using a Log-Likelihood Ratio (LLR) based speech activity detector [5]. The LLR of each frame is calculated between the speech and non-speech models with some predefined prior probabilities. To smooth LLR values, two adjacent windows with a same duration are located at the left and right sides of each frame and the average LLR is computed over each window. Thus, a frame is considered as a possible change point when a sign change is found between the left and right average LLR values. When several contiguous change candidates occur, the transition point is assigned to the maximum of difference between the averaged ratio of both windows. The SAD acoustic models, each with 256 Gaussians, were trained on about 2 hours of lecture data recorded at the University of Karlsruhe (UKA).
- Initial segmentation: the initial segmentation of the signal is performed by taking the maxima of a local Gaussian divergence measure between two adjacent sliding windows of 5 seconds, similar to the KL2 metric based segmentation [6]. Each window is modeled by a single diagonal Gaussian using the static features (i.e., only the 12 cepstral coefficients plus the energy).
- 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.

- 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 Viterbi resegmentation. The BIC criterion [7] is used both for the inter-cluster distance measure and the stop criterion. It is defined as:

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

  where d is the dimension of the feature vector space and λ weights the BIC penalty. At each step the two nearest clusters are merged, and the ΔBIC values between this new cluster and remaining clusters are computed. This clustering procedure stops when all ΔBIC are greater than zero. This BIC based clustering uses the size of the two clusters to be merged $n = n_i + n_j$ to calculate the penalty term, which is referred to as a local BIC measure. An alternative is to use the all frames in the whole set of clusters to compute the penalty, namely the global BIC penalty. Since the global penalty is constant, the clustering decision is made only by the likelihood increase. It has been experimentally found that the local BIC outperforms the global one on broadcast news data [8–10].
- 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 [14] of the means of an Universal Background Model [15]. This UBM is composed of 128 diagonal Gaussians and is trained on a few hours of the 1996/1997 English Broadcast News data. Then a second stage of agglomerative clustering is carried out on the segments within each gender class (male/female) separately according to the cross log-likelihood ratio as in [16]:

$$\mathcal{S}(c_i, c_j) = \frac{1}{n_i} log \frac{f(x_i|M_j)}{f(x_i|B)} + \frac{1}{n_j} log \frac{f(x_j|M_i)}{f(x_j|B)} \tag{2}$$

  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).

3 Applying voicing factor to SAD

Normally, the LLR-based speech activity detector is performed on cepstral coefficients with their Δ and $\Delta\Delta$ plus Δ and $\Delta\Delta$ log-energy. The reason for not using energy directly is that its level is sensitive to recording conditions, and it needs to be carefully normalized. Experiments on broadcast news data had led us to discard this feature. In order to further improve the SAD performance, a new energy normalization method taking into account a voicing factor v along with the energy E_0 is proposed. For each frame, the voicing factor is computed as the maximum peak of the autocorrelation function (excluding lag zero). The harmonic energy is thus defined as the energy associated with the best harmonic configuration, i.e. $E_h = v.E_0$. Finally, the energy of the signal is normalized relative to a reference level determined on the 10% frames carrying the highest harmonic energy. This way, the energy normalization focuses primarily on the voiced frames and may be more robust to varying SNR configurations. This method may be sensitive to music, but this is not expected to be an issue in the context of conferences and lectures.

4 Experimental results

In the RT-07S meeting recognition evaluation, LIMSI submitted speaker diarization results for both the conference and lecture meeting data on the MDM and SDM audio input conditions. For the MDM test condition, the speaker diarization system is performed on the beamformed signals created by the ICSI delay&sum signal enhancement system (see [2] for details) from all available input signals. Unless otherwise specified, the development experiments were carried out with a BIC penalty weight $\lambda = 3.5$ and a SID threshold $\delta = 0.5$.

Since the RT07S evaluation plans specified the use the word-forced alignment reference segmentations for scoring speaker diarization performance, the SAD acoustic models and UBM employed in the SID clustering stage are retrained using forced alignment segmentations. The forced alignment transcriptions were derived from the manual transcriptions via the ICSI-SRI ASR system [17] that aligns the spoken words to the signal, and therefore segment boundaries are labeled more accurately than the hand-made ones. Based on the aligned segmentations, more precise speech and non-speech models can be estimated and are expected to provide better SAD performances. Instead of training on the UKA lecture data that was used last year and in the baseline system, an union of several previous RT conference datasets consisting of recordings from different sites was used to better approximate this year's test data.

4.1 Performance measures

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 [18]. 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.

In addition to the above speaker diarization scoring metrics, the SAD measurement defined in the RT-06S evaluation [19] is used to measure the performance progress of the diarization system. The SAD task performance is evaluated by summing the missed and false alarm speech errors, without taking into account different reference speakers. The SAD error is normally included in the missed and false alarm speaker errors. Although the primary metric used in RT-07S evaluation is calculated over all the speech including the overlapping speech, the DER restricted to non-overlapping speech segments is also given for comparison purposes.

4.2 Corpus description

In order to tune the system parameters, the development experiments were carried out on the RT-06S test data in both the conference and lecture sub-domains. The conference development dataset *conf dev07s* is composed of 9 conference meetings, with a duration of about 15 minutes each. This is the same corpus that was used as the test data in the RT-06s evaluation and were provided by 5 different laboratories: CMU (Carnegie Mellon University), EDI (The University of Edinburgh), NIST, TNO (TNO Human Factor) and VT (Virginia Tech). The lecture development data includes two corpora: the RT-06S lecture evaluation dataset (denoted as *lect dev07s1*) and a new development dataset (denoted as *lect dev07s2*) released in 2007. The *lect dev07s1* consists of 38 5-minutes lecture excerpts contributed respectively by 5 of the CHIL partner sites: AIT, IBM, ITC, UKA and UPC, for which only 28 excerpts reference segmentations are available. The *lect dev07s2* contains 5 lectures with different audio lengths ranging from 23 minutes to 44 minutes, recorded more recently at the same 5 CHIL sites.

4.3 LLR-based SAD with different acoustic features

Although the speech activity detection task is not included in the RT-07S evaluation, a good SAD module is always useful for a speaker diarization system in the sense that it can influence the accuracy of the acoustic models which serve in the subsequent segmentation and clustering stages. Therefore, different kinds of acoustic features were investigated within the LLR-based SAD module. To do this, the SAD stage is separated from the speaker diarization system and assessed as an independent system on the RT-07S development data. However, an optimal SAD is not necessarily the best choice for diarization systems, as false alarm speech will corrupt the speaker models used in clustering stage: the experiments made at ICSI also show that minimizing short non-speech data is helpful to improve diarization performance [20].

Table 1 gives the SAD results with using different acoustic features, where each type of features has its appropriate SAD acoustic models estimated on the training data parameterized by the same features. In all cases, the speech and non-speech models are trained on the same data consisting of 8 RT-04S development conference meetings, 8 RT-04S evaluation conference meetings and 10 RT-05S evaluation conference meetings. It should be noted that the forced alignment segmentations provided by ICSI-SRI were used to estimate the SAD acoustic models. The lack of the forced alignments for the lecture data except the *lect dev07s1* (which serves as development data) is the reason of using only the conference meetings to construct the speech and non-speech models for both the conference and lecture test data.

The SAD results shown in Table 1 are obtained with the same LLR-based SAD configurations: the smoothing window with a size of 50 frames and the prior probabilities for the non-speech and speech models being $0.2 : 0.8$ with 256 Gaussians components in each model. The baseline vector consists of 12 cepstral coefficients with their Δ and $\Delta\Delta$ plus Δ and $\Delta\Delta$ log-energy and provides a SAD error of 5.6% and 7.8% on the conference and lecture development datasets respectively. When the raw energy is simply appended into the baseline acoustic vectors (c.f. denoted as "baseline+e" in Table 1), the SAD error is reduced to 5.1% for the conference dataset but increases to 11.5% for the lecture data. This degradation of the SAD performance on the lectures is predictable due to the variability of recording conditions in different lecture rooms. The SAD error reduction obtained on the conference meetings implies that the SNR configuration is consistent across the conference audio data. Replacing the energy with the normalized energy relying on the voicing factor v (denoted as "baseline+env") decreases the SAD error to 4.3% on the conference meeting data and 5.7% on the lecture meeting data. Regarding some details, the performance improvement obtained on the conferences comes from the reduction of the false alarm speech error, while the gain observed on the lectures derives merely from the missed speech error. The "baseline+e+mvn" feature set performs a variance normalization of each baseline feature and energy by subtracting their means and scaling by their standard deviations. Using this normalized acoustic representation, a further SAD reduction of absolutely 0.4% is obtained on the conference development data, but no improvement is obtained on the lecture dataset. For the simplicity of the diarization system, the SAD models trained with the "baseline+e+mvn" feature set were used for both the RT-07S conference and lecture evaluation data. Speech and non-speech models with 128 Gaussians were also investigated and although they gave a higher SAD error compared with 256 Gaussians, they perform slightly better on the diarization task.

Table 1. SAD results obtained with using different kinds of acoustic features on the RT-07S beamformed MDM development data.

acoustic features	*missed speech error (%)*	*false alarm speech error (%)*	*overlap SAD error (%)*
conf dev07s			
baseline	1.3	4.3	5.6
baseline+e	1.1	4.0	5.1
baseline+env	1.1	3.3	4.3
baseline+e+mvn	0.8	3.0	3.9
lect dev07s1			
baseline	2.4	5.3	7.8
baseline+e	0.5	11.2	11.8
baseline+env	0.9	4.7	5.7
baseline+e+mvn	1.0	5.6	6.6

4.4 SID clustering with UBMs trained on different acoustic features

The different sorts of acoustic features are also tested for the UBM training. A single gender-independent UBM with 128 Gaussian mixtures is trained for each type of feature, since no gender information is available in the forced alignment segmentations of the training data. Table 2 shows the speaker diarization results on the beamformed MDM development data using different acoustic features to train UBMs. These results are obtained with the same SAD acoustic models that were trained on the normalized features via the variance normalization technique. All UBMs are trained on the same conference dataset that have been used to estimate the speech and non-speech models.

Table 2. Diarization results obtained with the UBMs trained on different acoustic representations on the RT-07S beamformed MDM development data (results with '*' were obtained after the evaluation).

acoustic features	*speaker match error (%)*	*overlap DER (%)*
conf dev07s		
15plp+w*	20.5	28.3
15plp+Δ+ΔlogE+w	28.4	36.2
15plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	23.3	31.1
12plp+w*	21.3	29.0
12plp+Δ+ΔlogE+w	22.9	30.6
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	27.9	35.7
12plp+mvn*	28.6	36.3
12plp+Δ+ΔlogE+mvn	33.8	41.6
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+mvn	32.0	39.8
lect dev07s1		
15plp+w*	10.3	17.8
15plp+Δ+ΔlogE+w	10.0	17.5
15plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	10.2	17.7
12plp+w*	11.3	18.8
12plp+Δ+ΔlogE+w	10.3	17.8
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	10.2	17.7
12plp+mvn*	10.1	17.6
12plp+Δ+ΔlogE+mvn	10.5	18.0
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+mvn	10.2	17.7

The baseline feature vector is composed of 15 Mel frequency cepstral coefficients plus the Δ coefficients and the Δ log-energy with the feature warping normalization (referred to as "15plp+Δ+ΔlogE+w" in Table 2). This baseline feature set provides an overlap DER of 36.2% on the conference development data and 17.5% on the lecture development data. Adding the ΔΔ coefficients and the ΔΔ log-energy, namely "15plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w" reduces the DER to 31.1% on the conferences but gives a similar diarization performance as the baseline features for the lectures. When the dimension of cepstral coefficients is diminished to 12, using the "12plpΔ+ΔlogE+w" feature set results in a further DER reduction of 0.5% on the conference data but no significant performance change on the lecture meetings. Appending the ΔΔ coefficients (denoted as "12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w"), a large increase of DER is observed on the conference data but the DER score rests always very close to the baseline one for the lectures. Finally, we also examined the variance normalization method within the SID clustering stage. As can be seen in Table 2, higher DER rates are provided by this normalization approach than the feature warping technique on both the conference and the lecture development data. For the lecture data, the different acoustic representations are found to give similar diarization results. This may be derived from the mismatch between the conference training data and the lecture test data.

The results of the post-evaluation experiments without the use of derivative parameters are also given in Table 2[1]. Using only static features reduces the diarization error rates on the conference development data and the lowest DER of 28.3% is given by the 15 MFCC with the feature warping normalization. However, no significant difference in diarization performances is observed on the lecture data between using only the static coefficients and by appending the Δ and $\Delta\Delta$ coefficients.

4.5 RT-07S evaluation results

The speaker diarization results for the systems submitted to the RT-07S evaluation are given in Table 3. The diarization system uses the same SAD acoustic models and UBM trained on the "baseline+e+mvn" and the "12plp+Δ+ΔlogE+w" feature sets respectively for both the conference and the lecture evaluation data. The BIC penalty weight λ and the SID clustering threshold δ were optimized on the development data and set individually for the conference and lecture test data: $\lambda = 3.5, \delta = 0.6$ for the conference dataset and $\lambda = 3.5, \delta = 0.5$ for the lecture dataset. For each type of the data, the same system configurations were used on the MDM and SDM audio input conditions.

For the conference test data, the diarization system has an overall diarization error of 26.1% on the beamformed MDM signals, and the overall DER increases to 29.5% for the SDM condition. The beamformed signals from all available distant microphones are shown to be helpful for improving the diarization performance. For the lecture evaluation data, the diarization system gives similar performances on both the beamformed MDM signals and a single SDM data. This may be because the delay&sum signal enhancement system was not well tuned for lecture data.

Table 3. Speaker diarization performances on the RT-07S conference and lecture evaluation data for the MDM and SDM conditions.

data type & condition	*missed speaker error (%)*	*false alarm speaker error (%)*	*speaker match error (%)*	*overlap DER (%)*	*non-overlap DER (%)*
conference MDM	4.5	1.3	20.2	**26.1**	23.0
conference SDM	4.9	1.3	23.3	**29.5**	26.6
lecture MDM	2.6	8.4	14.7	**25.8**	24.5
lecture SDM	2.9	8.1	14.7	**25.6**	24.3

[1] We thank the reviewers for suggesting this contrastive experiment.

5 Conclusions

The LIMSI speaker diarization system for meetings within the framework of the RT-07S meeting recognition evaluation was described in this paper. The RT-07S diarization system for both conference and lecture meetings keeps the main structure of the RT-06S lecture diarization system except removing the bandwidth detection module, since it is supposed that no telephonic speech would occur during the meetings. The main improvements come from the new SAD acoustic models and UBM that were built on the conference training data with their forced alignment segmentations. Using the speech and non-speech models trained with the variance normalized acoustic features yields the best SAD performance on the development data, i.e. 3.9% for the conference data and 6.6% for the lecture data. The mismatch between the conference training data and the lecture test data results in a relatively higher SAD error on the lecture development data. As for UBMs, the best diarization performance is generated by using the gender-independent UBM trained with 12 Mel frequency cepstral coefficients plus Δ coefficients and Δ log-energy with the feature warping technique. The adapted diarization system provides similar diarization results on the beamformed MDM signals for both the RT-07S conference and lecture evaluation data (i.e. an overlap DER of 26.1% for the conference dataset and 25.8% for the lecture dataset). The DER rate increases to 29.5% on the conference SDM data, while for the lecture SDM data, the error rate remains very closely to the one obtained on the beamformed MDM condition.

References

1. NIST, “Spring 2007 Rich Transcription (RT-07S) Meeting Recognition Evaluation Plan,” `http://www.nist.gov/speech/tests/rt/rt2007/spring/docs/rt07s-meeting-eval-plan-v2.pdf`, February, 2007.
2. X. Anguera, C. Wooters, and J. Hernando, “Speaker Diarization for Multi-Party Meetings Using Acoustic Fusion”, in *Automatic Speech Recognition and Understanding (IEEE, ASRU’05)*, San Juan, Puerto Rico, 2005.
3. 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.
4. 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.
5. X. Zhu, C. Barras, L. Lamel and J.L. Gauvain, “Speaker Diarization: from Broadcast News to Lectures”, In *MLMI 2005 Meeting Recognition Workshop*, Washington DC, USA, May 2006.
6. 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.
7. 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.
8. 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.

9. C. Barras, X. Zhu, S. Meignier and J.L. Gauvain, “Improving speaker diarization” in *the Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, USA, November 2004.
10. S. E. Tranter and D. A. Reynolds, “Speaker diarization for broadcast news,” in *Proc. ISCA Speaker Recognition Workshop Odyssey 2004*, Toledo, Spain, May 2004.
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*, June 2001, pp. 213–218.
14. 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.
15. 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.
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. A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters and J. Zheng, “The ICSI-SRI Spring 2005 Speech-To-Text evaluation System,” in *Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Edinburgh, Great Britain, July, 2005.
18. 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.
19. 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.
20. C. Wooters and M. Huijbregts, “The ICSI RT07s Speaker Diarization System”, In *Rich Transcription 2007 Meeting Recognition Workshop*, Baltimore, USA, May 2007.