IMPROVEMENTS IN SPEAKER DIARIZATION SYSTEM

Rong Fu and Ian D. Benest
Department of Computer Science, the University of York, YO10 5DD, York, UK

Keywords: Speaker Diarization, Model Complexity Selection, Universal Background Model.

Abstract: This paper describes an automatic speaker diarization system for natural, multi-speaker meeting conversations using one central microphone. It is based on the ICSI-SRI Fall 2004 diarization system (Wooters et al., 2004), but it has a number of significant modifications. The new system is robust to different acoustic environments - it requires neither pre-training models nor development sets to initialize the parameters. It determines the model complexity automatically. It adapts the segment model from a Universal Background Model (UBM), and uses the cross-likelihood ratio (CLR) instead of the Bayesian Information Criterion (BIC) for merging. Finally it uses an intra-cluster/inter-cluster ratio as the stopping criterion. Altogether this reduces the speaker diarization error rate from 25.36% to 21.37% compared to the baseline system (Wooters et al., 2004).

1 INTRODUCTION

For the purposes of this paper, speaker diarization is the process by which an audio recording of a meeting is indexed according to the speakers who made oral contributions. The NIST Rich Transcription Evaluations (NIST, 2004) refers to this as “who spoke when”. Research in speaker diarization currently focuses on three main areas. First, through its indexing and modelling, diarization enables audio databases to be searched for particular individuals. Second, the provision of diarization improves the success rate of automatic speech recognisers by enabling them to adapt to different speakers. The third application is in the provision of more structured transcripts of recorded meetings, news broadcasts and telephone conversations (Tranter and Reynolds, 2006). It is the first area on which this paper focuses. The process reported here is performed without the knowledge of, for example, the number of speakers involved, their gender, and the positions of extraneous noises such as laughter, coughing, paper shuffling and so on. While meetings can be recorded using either a single central microphone or multiple-microphones (where each person has their own microphone) (Jin et al., 2004), this work concentrates on single-microphone recordings of meetings between a number of individuals. Of course the system needs to be robust with regard to the acoustic environment - implying that there should be no pre-training of the acoustic models, and the tuning of parameters should be automatic.

The system described in this paper is based on the ICSI-SRI Fall 2004 diarization system (Wooters et al., 2004). This was selected because it adopts the single microphone speaker diarization task, and performs acoustic modelling using only the audio file itself. In contrast with their system (Wooters et al., 2004), this new one adopts an alternative approach to determining the model complexity parameters automatically, using the cross log-likelihood ratio rather than the Bayesian Information Criterion (BIC) as the merging rule; in addition it uses the intra-cluster/inter-cluster ratio as the criterion for identifying the number of speakers.

The paper is structured as follows. Section 2 describes the basic diarization system adopted by Wooters et al.(2004), which is taken to be the baseline for comparison. In section 3 the techniques used to improve on the basic system are described. The experimental arrangement and results are presented in section 4 and conclusions are offered in section 5.

Fu R. and D. Benest I. (2007).
IMPROVEMENTS IN SPEAKER DIARIZATION SYSTEM.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 313-319
DOI: 10.5220/0002140703130319

2 THE BASELINE DIARIZATION SYSTEM

In the ICSI-SRI Fall 2004 diarization system a guess is made as to the number of individual speakers (K); that guess must be much greater than the number of actual speakers. The audio file is divided up into 60 millisecond windows with each window overlapping the previous one by 20 milliseconds. For each window, nineteen mel-frequency cepstral coefficients (MFCC) are extracted as acoustic feature vectors for that window. These feature vectors are assigned sequentially to the K speakers; this grouping of feature vectors is called a segment (and there are K segments). Wooters et al. (2004) report that this speaker change detection initialization method is as effective as those based on distance measures (Barras et al., 2004) or BIC (Zhou and Hansen, 2000).

A K state Hidden Markov Model (HMM) is created where, of course, each of its states acoustically models a single potential speaker. Gaussian Mixture Models (GMM) are established to initialize the states of the HMM. The Viterbi decoding algorithm is used to re-assign feature vectors to other states and the GMM is thus updated. Several sub-states are linked to each K state and these share the state's probability density function (pdf). Upon entering a state, the feature vectors cannot change to another state unless they have travelled through all the sub-states one-by-one. This imposes a minimum number of features (equivalent to more than 0.9 seconds), which are assigned to a state each time. This iteratively refines the segment boundary assigned to each state. This approach was first reported by Ajmera et al.(2002).

Wooters et al. (2004) advise that an agglomerative clustering technique with BIC merging and stopping criteria (Ajmera and Lapidot, 2002) always gives the best performance for clustering segments. Bayesian Information Criterion (BIC) (Schwarz, 1978) is a model selection criterion which prefers those models that have large log-likelihood values, but penalizes it with model complexity (the number of parameters in the model) (Schwarz, 1978). For a pair of segments x and y which are assigned to different states, their BIC merging score is computed according to Eq.1.

$$BIC_{score} = L_z - (L_x + L_y) - 1/2\alpha(P_z \\ log(n_z) - P_x log(n_x) - P_y log(n_y)), \qquad (1)$$

where L_z is the log-likelihood function for the merging model, P is the number of parameters used in the model and n is the number of features in the segment. The pair of states whose segments have the highest BIC score will be merged, and the state model retrained. The merging process continues until there are no pairs of states whose BIC score is larger than zero; the clustering then stops. In the ICSI-SRI diarization system, the number of parameters used in the merging model is set to be equal to the sum of the number of parameters used in each model, so the α parameter is not required. The states that remain in the HMM are potential speakers; the segments are thus indexed and categorised. Ajmera and Wooters (2003) have created an alternative algorithm which integrates the segmentation and clustering together.

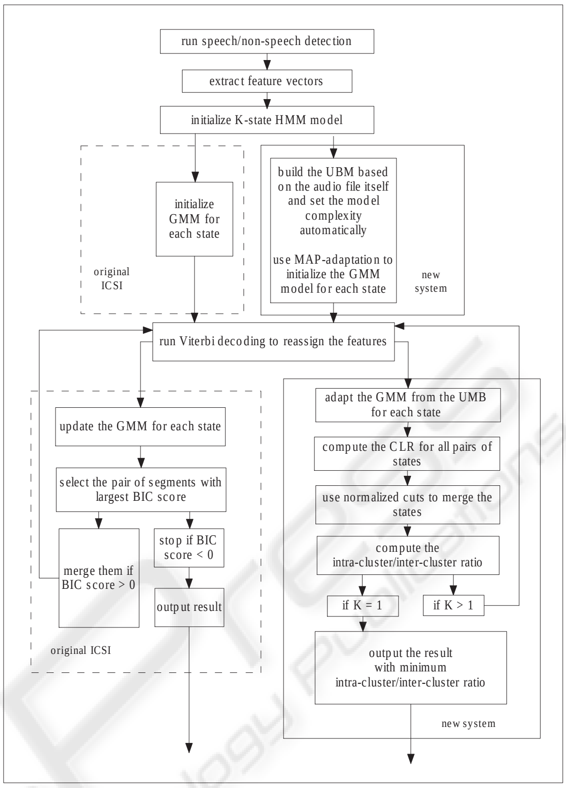
run speech/non-speech detection
extract feature vectors
initialize K-state HMM model
initialize GMM for each state
original ICSI
build the UBM based on the audio file itself and set the model complexity automatically
use MAP-adaptation to initialize the GMM model for each state
new system
run Viterbi decoding to reassign the features
update the GMM for each state
select the pair of segments with largest BIC score
merge them if BIC score > 0
stop if BIC score < 0
output result
original ICSI
adapt the GMM from the UMB for each state
compute the CLR for all pairs of states
use normalized cuts to merge the states
compute the intra-cluster/inter-cluster ratio
if K = 1
if K > 1
output the result with minimum intra-cluster/inter-cluster ratio
new system

Figure 1: The original ICSI method compared with the new system.

Sinha and Tranter (2005) and Barras et al. (2006) have included a post-processing step in the speaker diarization system in order to improve the performance. This involves a Universal Background Model (UBM), which is pre-trained either with other audio files or with the data itself, and a Maximum a Posteriori (MAP) mean-adaptation (Barras and Gauvain, 2003) is then applied to each cluster from the UBM to give the state model. The Cross-Likelihood Ratio (CLR) (Sinha et al., 2005) instead of BIC is applied as the merging criterion.

Figure 1 illustrates and contrasts the original ICSI-SRI system with that described by the authors.

3 THE NEW APPROACH

3.1 Model Complexity Selection

Model complexity is defined as the number of parameters within the model. For the Gaussian Mixture Model with a fixed dimension and covariance type (diagonal or full), the parameter which determines the model complexity is the number of components in the mixture model. In speaker diarization systems, there are two steps which are influenced by the model complexity; which pair of segments are to be merged, and how should the model for a new segment be established. A GMM with a small number of components is too general and tends to over-merge the segments, while a GMM with a large number of components is too specific and tends to under-merge the segments. Usually GMMs are trained by the EM algorithm and unfortunately this is sensitive to the initialization environment. As a result an inappropriate estimate of the number of components reduces the accuracy of the model.

Anguera et al. (2005), and Wooters et al. (2004) pre-determine this complexity with a fixed value; so the model complexity cannot be changed according to segment length and model similarity. The problem is that a small segment with high complexity loses its generalization. Anguera et al. (2006) automatically determined the number of components depending on the length of the segments. However, there was still a need for a parameter that was adjusted by external training sets. A problem also arose from the merging process: the model of a populated segment doubles its size after merging two segments of the same length, ignoring both the similarity between the two segments and their common speaker characteristics. The problem becomes more serious when UBM is introduced to derive the segment models - as discussed in section 3.2.

The approach developed by Figueiredo and Jain (2002), which overcomes the sensitivity limitation of the EM algorithm, was adopted in the new system, but with a modified stopping criterion. It automatically determines the model complexity, and gives a model that better fits the data.

Consider a finite data set $X = \{x^1, x^2, \dots x^n\} \subset \Re^d$, and an M-component GMM finite mixture distribution of dataset X, its pdf can be written as:

$$p(x|\theta) = \sum_{m=1}^{M} \pi_m p(x|\mu_m, \sigma_m). \qquad (2)$$

where $\sum_{m=1}^{M} \pi_m = 1$ and $\pi_m \leq 1, m = 1, \cdots, M$. π_m is the mixing parameter of the mixture model, and μ_m and σ_m are the mean and covariance parameters of the Gaussian component m. Then $\log p(X|\theta)$ is defined as:

$$\log p(X|\theta) = \sum_{i=1}^{n} \log \sum_{m=1}^{M} \pi_m p(x^{(i)}|\mu_m, \sigma_m). \qquad (3)$$

$$\theta = \{\pi, \mu, \sigma\} \qquad (4)$$

Usually an EM algorithm is applied to obtain the maximum likelihood (ML) estimate $\hat{\theta}_{ML}$:

$$\hat{\theta}_{ML} = \text{argmax}_{\theta}\{log p(X|\theta)\} \qquad (5)$$

The EM algorithm (McLachlan and Krishnan, 1997) runs iteratively with an E-step followed by an M-step. The E-step computes the conditional expectation of the complete log-likelihood, given the dataset and current estimate of all parameters - the so-called Q-function. The M-step updates the estimate of θ in order for it to maximize the Q-function (McLachlan and Krishnan, 1997). The EM algorithm will monotonically increase the $\log p(x|\theta)$ value until it reaches one of the local maxima. Using the EM algorithm to train the GMM model, label variables $Z = \{z^1, \cdots, z^n\}$ are introduced so as to indicate which component produces which sample. The conditional expectations of z^i_m, where z^i_m is the label which shows whether data x_i is produced by component m at iteration t, is given by:

$$o^t_m(x_i) = E[z^t_m|x_i, \hat{\theta}^t] = \frac{\hat{\pi}^t_m p(x^i|\hat{\theta}^t_m)}{\sum_{j=1}^{M} \hat{\pi}^t_m p(x^i|\hat{\theta}^t_m)} \qquad (6)$$

and π^{t+1}_m will be updated thus:

$$\pi^t_m = \sum_{i=1}^{n} o^t_m(x_i)/n; \qquad (7)$$

o^i_m is the posteriori probability of Z^i_m, given the observation x^i. Since the EM algorithm theoretically searches the local maxima, its result is sensitive to its initialization values. The results can always be adjusted to get a better global maximum by splitting or merging the components (Ueda et al., 2000). Figueiredo and Jain (2002) developed an algorithm that successfully determines the number of components used in building the GMM, and, at the same time, optimizes the distribution of components. It seamlessly integrates the Minimum Message Length (MML) criterion into the EM algorithm. MML, like BIC, is a model selection criterion. It is based on information coding theory and was first developed by (Wallace and Dowe, 1987). It prefers the model which minimizes the right-hand side of Eq.8:

$$\begin{aligned} Length(\hat{\theta}, X) &= -\log p(\hat{\theta}) - \log p(X|\hat{\theta}) \\ &+ \frac{1}{2}\log|I(\hat{\theta})| + p/2(1 + \log(1/12)), \end{aligned} \qquad (8)$$

where p is the number of parameters in the model. Figueiredo and Jain (2002) assumed an independence between the mixing parameter π and the component parameter $\hat{\theta}(m)$, and expresses a standard non-informative Jeffreys's prior for π, and $\hat{\theta}$,

$$p(\hat{\theta}_m) \propto \sqrt{|I^{(1)}(\hat{\theta}_m)|}; \qquad (9)$$

$$p(\pi_1,\cdots,\pi_M) = \sqrt{|H|} = (\pi_1\pi2\cdots\pi_M)^{-1/2}; \quad (10)$$

and approximates $|I(\theta)|$ to $|I_c(\theta)|$, the complete-data Fisher information matrix, which is described by Eq.11:

$$I_c(\theta) = n * block - diag\{\pi_1 I^1(\hat{\theta}_1),\cdots,\pi_M I^1(\hat{\theta}_M), H\}, \qquad (11)$$

where $I^1(\theta_m)$ is the Fisher matrix for a single observation produced by the mth component, with $|H|$ defined in Eq.10. Then Eq.8 can be rewritten as:

$$Length(\hat{\theta},X) = \frac{N}{2}\sum_{m=1}^{M}\log(n\pi_m) + \frac{p}{2}(1+\log\frac{1}{12}) + \frac{p}{2}\log(n), \qquad (12)$$

where p in Eq.8 is equal to $NM+M$, where N is the number of parameters in each component. When the π_m of a component m is equal to 0, the component will no longer contribute to the model, so then M is the number of components whose mixing parameter π_m is larger than 0. Eq.12 is the criterion provided by Figueiredo and Jain (2002), which can be integrated into the EM algorithm in a closed form. In Eq.12, the term $\frac{N}{2}\sum_{m=1}^{M}\log(n\pi_m)$ containing the variable π is combined with Eq.5; then this is maximized:

$$\frac{\partial(\log p(X|\hat{\theta}))}{\partial\pi_m} + \lambda(\sum_{m=1}^{M}\pi_m - 1) + \frac{N}{2}\sum_{m=1}^{M}\log(n\pi_m)) = 0, \qquad (13)$$

Instead of Eq.7, Eq.14 is calculated as the update value for π_m:

$$\pi_m^t = \frac{1}{n-MN/2}(\sum_{i=1}^{n}\max(o_m^t(x^i) - N/2, 0)) \qquad (14)$$

where $o_m^t(x^i)$ is defined as in Eq.6.

By initializing the model complexity to a large value M, this algorithm will reduce the complexity value to M_n by removing the components that have insufficient evidence to support them. But there may be an additional decrease in Eq.12 caused by a decrease in M_n. So the algorithm needs to compute the criterion value for each possible M_n in order to get the best result. This is computationally expensive when the number of components is large. An alternative stopping criterion is the local Kullback-Leibler divergence criterion in which those components with the lowest local Kullback-Leibler divergence are removed and the model retrained until the value of the right side of Eq.12 no longer decreases. The local Kullback-Leibler divergence criterion of component m is described by Eq.15:

$$J_{merge}(m|\hat{\theta}) = \int f_m(x|\hat{\theta})\log\frac{f_m(x|\hat{\theta})}{f'_m(x|\hat{\theta})}dx, \qquad (15)$$

where f_m is the current density function g, and f'_m is the density function g', whose component m has been removed, all weighted by the posteriori function of m:

$$f_m(x|\hat{\theta}) = g(x_i)p(m|x_i,\hat{\theta}); \qquad (16)$$

$$f'_m(x|\hat{\theta}) = g'(x_i)p'(m|x_i,\hat{\theta}); \qquad (17)$$

3.2 UBM and MAP Adaptation

Sometimes the segment is short and may be disturbed by the acoustic environment so that the model on which it is based is not sufficient to represent the speaker characteristics. Then the Universal Background Model (UBM) is introduced to derive the segment models as referred to earlier in section 2.

UBM is always pre-trained from other speech corpora. The audio file itself can be used to train the UBM, or a combination of corpora and the speech file may be used (Sinha et al., 2005) (Barras et al., 2006). The authors adopted the audio file itself to build the UBM and used the technique described in the last section to develop the background model. A mean-only MAP adaptation is always used to build the segment model from the UBM. It updates the components' means in the segment model by adapting them from the UBM gradually. However, sometimes the segments produced by the system are very short and not sufficient to cover the space modelled by the UBM. As a result there may be some loss of segment characteristics - hence the reason why this technique is always applied as a secondary stage.

The authors' system adjusts the weight parameter in the UBM and removes those components for which there is insufficient evidence (according to Eq.14) for their retention. This step is used at the beginning of the process, so determining the complexity of the UBM model is important. The adapted weight estimator is described by Eq.18

$$\hat{\pi}_m = \frac{\max(\sum_{i=1}^{n}p(m|x^i) - N/2, 0)}{\sum_{m=1}^{k}\max(\sum_{i=1}^{n}p(m|x^i) - N/2, 0)}, \qquad (18)$$

where N is the number of parameters in each component, and $p(m|x^i)$ is the posteriori probability of component m given x_i. Usually the mean-MAP adaptation is then performed (Barras and Gauvain, 2003):

$$\hat{\mu}_m = \frac{\frac{1}{2}\mu_i + \frac{1}{2}(\sum_{i=1}^{n} p(m|x_i)x_i)}{\frac{1}{2} + \frac{1}{2}\sum_{i=1}^{n} p(m|x_i)}, \qquad (19)$$

This adaptation is performed for two iterations.

3.3 Cross-likelihood Ratio

CLR is used in place of BIC as the merging measure adopted in most post-processing stages. Between any two given segments s_i and s_j, CLR is defined by:

$$CLR(s_i, s_j) = \log(\frac{L(x_i|\theta_j)L(x_j|\theta_i)}{L(x_i|\theta_{ubm})L(x_j|\theta_{ubm})}), \qquad (20)$$

where $L(x_i|\theta_j)$ is the average likelihood of the acoustic feature being in segment i given the model j, thus removing the influence of the length of the segments (Sinha et al., 2005). The CLR values are computed for each pair of segments to form the CLR matrix.

3.4 Intra-cluster/Inter-cluster Ratio

Using BIC to merge and judge the stopping of the speaker diarization task is, computationally, a local solution. It considers those pairs of segments that have the highest BIC score, without a global view of the overall similarity between segments. The authors found that it always under-estimated the number of speakers appearing in the audio file. Futhermore in the new system, CLR is applied as the merging criterion instead of BIC; and BIC is no longer the naturally stopping criterion. So in place of BIC, an intra-cluster/inter-cluster ratio was used as the stopping criterion. Intra-cluster/inter-cluster ratio compares the way in which a state model represents its features, with how the models of other states represent those features; in so doing, it takes a global view of all the states. Assuming that there are k clusters left in the HMM, the intra-cluster/inter-cluster ratio is computed by Eq.21:

$$\frac{intra-cluster}{inter-cluster} = \sum_{i=1}^{k} \frac{CLR(s_i, s_i)}{\sum_{j,j\neq i} CLR(s_i, s_j)}. \qquad (21)$$

The new system iteratively merges the segments until there is only one segment left. Then the speaker diarization solution is the cluster solution with minimum intra-cluster/inter-cluster ratio.

4 EXPERIMENT

4.1 Data

The experiments reported in this paper used 36 recorded meetings as evaluation data, 18 from the Interactive Systems Laboratories (ISL) part A meeting corpus (ISL, 2004) and 18 from the International Computer Science Institute (ICSI) meeting speech corpus (ICSI, 2004). The average meeting length was around 40 minutes, and the number of speakers present in the meetings varied from 3 to 9. In the meeting audio files, speaker changes are frequent, and most have a short duration. Many exchanges overlap with each other, hindering the speaker diarization.

4.2 Speaker Diarization Error

The baseline metric of performance is the diarization error rate (DER), which is defined by NIST (2004). It is obtained by comparing the results of diarization, with a manually labelled transcript. The DER process recognises three error types: the missed rate (speech in the transcript that is not found by the system under test), false alarm rate (speech found by the system that is not in the transcript) and speaker error rate (the sound is assigned to the wrong speaker). These are computed by matching the speakers assigned by the system to those in the transcript, using a one-to-one mapping that maximises the total overlap between the transcript and system speakers (as explained in (NIST, 2004)).

The main application of the new system is to index and cluster the audio files. Usually the type of non-speech appearing in the meeting audio files is silence and noise. The false alarm rate does not influence the indexing and clustering, nor does the missed speech fragments as these last for less than 0.3 seconds. So an energy based speech/non-speech detection is enough. Since DER is a time-weighted evaluation measure, it is primarily driven by dominant speakers. To reduce the DER error rate, it is more important to find the main speakers completely and correctly rather than accurately find the speakers who do not speak very often. Therefore when there is a dominant speaker in the audio file, the DER measure fails to evaluate how well the sytem finds other speakers. So segment purity (SEGP) and speaker weighted speaker purity (SPKP) are also introduced to evaluate the system. Segment purity is a measure of how much speech in a segment belongs to the correct speaker. Speaker purity measures how much speech is assigned correctly to a speaker, weighted by the number of speakers.

In the new system, the 19^{th} order MFCC was used as acoustic feature vectors, extracted from 30 millisecond analysis windows each overlapping the previous one by 20 milliseconds. The initial guess at the number of speakers was 40, and the minimum length constraint was 0.9 seconds.

4.3 Conversation Overlap

Since there are many oral exchanges that overlap each other in the audio files, it is possible that these reduce system performance, and this was investigated in three experiments. First, the overlapping parts were removed from the audio file and the diarization system applied. Second, the overlapping parts were included in the speaker diarization process, but ignored in the evaluation process. Third, the overlapping parts were included in the speaker diarization process and, during evaluation, those parts were assigned to a speaker. Perhaps surprisingly, both the second and the third experiments gave better results than the first experiment. It seems that the overlapping parts actually help to build the speaker models and thus do not degrade the recognition rate. However, sometimes the overlapped parts are regarded as new speakers, making the number of people detected more than those actually present. The third approach gives the best result because the overlapping speech assigned to either of the speakers is thought to be correct.

4.4 Model Complexity

Three different methods for determining the model complexity were investigated: first, fixing the number of Gaussians in the GMM; second, automatically determining the value for model complexity depending on the segment length - the method adopted by Anguera et al. (2006); and third, doing the same as number two, but using the approach introduced for the new system. A diagonal GMM was used and overlapped speech was ignored during the evaluation. All three were integrated into the original ICSI system, and their results are compared in Table 1. The approach used in the new system performed well when the segments were short. However, it is computationally expensive to find the number of components in a long segment, and only gives slightly better results. Since the approach is robust to the initialization environment, its results are stable.

Table 1: Model-Complexity-Determining Methods.

method	DER(%)	SEGP(%)	SPKP(%)
5 Gaussian/GMM	25.36	26.83	32.76
8 Gaussian/GMM	26.35	28.76	34.76
Anguera et al	22.25	22.76	28.76
authors' approach	22.74	24.76	30.76

4.5 Applying the UBM

This part provides the biggest improvement. The new EM approach used to train the GMM can automatically remove the inefficient components, helping to adapt the segment model from the UBM, especially for short segments. Table 2 shows the difference between the segment model adapted from the UBM, and that obtained from the ICSI system. Again, the overlapped speech was ignored during the evaluation. The UBM was built from the audio data itself and its model complexity was automatically determined. The segment models were then adapted from the UBM by weight and mean-MAP. BIC was retained for determining the merging and stopping criteria.

Table 2: Effect of UBM-MAP Adaptation.

method	DER(%)	SEGP(%)	SPKP(%)
UBM-MAP	19.74	21.85	28.36
Non UBM-MAP	25.36	26.83	32.76

4.6 Merging and Stopping Criteria

Finally, CLR and intra-cluster/inter-cluster ratio were used instead of BIC for the merging and stopping criteria. CLR seems to work well with the UBM adapted segment model. It considers the similarity between the segments and their similarity with the UBM. The intra-cluster/inter-cluster stopping criterion still under-estimates the number of speakers; this may happen when there is a dominant speaker at the meeting or the utterances from some speakers are short. This stopping criterion almost correctly detected the number of speakers, but the overall speaker error rate was degraded. As shown in Table 3, the new system reduces the error rate from 25.36% to 21.37%, an improvement of about 19%. These results are obtained when the overlapped parts were ignored during

Table 3: Overall Performance (without the overlap speech).

method	DER(%)	SEGP(%)	SPKP(%)
Baseline System	25.36	26.83	32.76
New System	21.37	22.56	27.40

the evaluation. The best results, for both the baseline system and the new one, were achieved when the results were evaluated using the approach that included the overlapped parts, as shown in Table 4.

Table 4: Overall Performance (including the overlap).

method	DER(%)	SEGP(%)	SPKP(%)
Baseline System	21.76	23.23	29.76
New System	17.21	18.56	26.40

In comparison with the ICSI system the diarization error rate was reduced from 21.76% to 17.21%.

5 CONCLUSION

This paper has described a new speaker diarization system for natural, multi-speaker meeting conversations based on a single microphone. The system requires no prior training, and the experiments showed that the performance achieved 21.37% speaker diarization error rate. It is expected that different speech/non-speech detection techniques and further purity tests will improve the performance of the system.

REFERENCES

Ajmera, J. and Lapidot, I. (2002). Improved unknown-multiple speaker clustering using hmm. In *IDIAP RR.* pp.02–23.

Barras, C. and Gauvain, J. L. (2003). Feature and score normalization for speaker verification of cellular data. In *ICASSP Proc.*

Barras, C., Zhu, X., Meignier, S., and Gauvain, J. L. (2004). Improving speaker diarization. In *Fall 2004 Rich transcription Workshop (RT-04) Proc.*

Barras, C., Zhu, X., Meignier, S., and Gauvain, J. L. (2006). Multistage speaker diarization of broadcast news. In *IEEE Trans. SL Proc.* pp.1505–1512.

ICSI (2004). Icsi meeting speech. In *International Computer Science Institute*. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp? catalogId=LDC2004S02.

ISL (2004). Isl meeting speech part 1, 2004. In *Interactive Systems Laboratories*. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp? catalogId=LDC2004S05.

Jin, Q., Laskowski, K., Schultz, T., and Waibel, A. (2004). Speaker segmentation and clustering in meetings. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Meeting Recognition Workshop Proc.*

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. John Wiley & Sons, New York, 1st edition.

NIST (2004). Fall 2004 rich transcription (rt-04) evaluation plan, 2004. In *National Institute of Standards and Technology*. Available:http://www.nist.gov/speech/tests/rt/rt04 /fall/docs/rt04f-eval-plan-v14.pdf.

Schwarz, G. (1978). Estimating the dimension of a model. In *Annals of Statistics Proc.* Vol.6, pp.461–464.

Sinha, R., Tranter, S. E., Gales, M. J., and Woodland, P. C. (2005). The cambridge university march 2005 speaker diarization system. In *Eur. Conf. Speech Communication Technology, Proc.* pp.2437–2440.

Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. In *IEEE Trans. Speech and Language (SL) Proc.* Vol.14, pp.1557–1565.

Ueda, N., Nakano, R., Gharhamani, Z., and Hinton, G. (2000). Smem algorithm for mixture models. In *Neural Computation Proc.* Vol.12, pp.2109–2128.

Wallace, C. and Dowe, D. (1987). Estimation and inference via compact coding. In *J. Royal Statistical Soc. (B)*. Vol.49, pp.241–252.

Wooters, C., Fung, J., Peskin, B., and Anguera, X. (2004). Toward robust speaker segmentation: Icsi-sri fall 2004 diarization system. In *Fall 2004 Rich transcription Workshop (RT-04) Proc.*

Zhou, B. and Hansen, J. (2000). Improving speaker diarization. In *Int. Conf. Spoken Langrage Process Proc.* Vol.3, pp.714–717.