THE DEVELOPMENT OF THE CAMBRIDGE UNIVERSITY RT-04 DIARISATION SYSTEM

S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, P. C. Woodland

Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK.
Email: {sej28,mjfg,rs460,su216,pcw}@eng.cam.ac.uk

ABSTRACT

This paper describes the development of the Cambridge University RT-04 diarisation system, including details of the new segmentation and clustering components. The final system gives a diarisation error rate of 23.9% on the RT-04 evaluation data, a 34% relative improvement over the RT-03s evaluation system. A further reduction down to 18.1% is shown to be possible when using the segmentation algorithm alone.

1. INTRODUCTION

Speaker diarisation is the task of automatically segmenting audio data and providing speaker labels for the resulting regions of audio. This has many applications such as enabling speakers to be tracked through debates, allowing speaker-based indexing of databases, aiding speaker adaptation in speech recognition and improving readability of transcripts.

The Rich Transcription diarisation evaluations[1, 2, 3] provide a framework to analyse the performance of such speaker diarisation systems on Broadcast News (BN) data. A Diarisation Error Rate (DER) is defined which considers the sum of the missed, false alarm and speaker-error rates after an optimal one-to-one mapping of reference and hypothesis speakers has been performed. (This mapping is necessary to associate the 'relative' speaker labels such as 'spkr1' from the hypothesis to the 'true' speaker labels such as 'Ted Koppel' in the reference).

Cambridge University first built a complete diarisation system in late 2002 and has participated in the diarisation evaluations since then. This paper describes the development of the Cambridge University diarisation system used in the Fall 2004 Rich Transcription evaluation (RT-04)[3, 4].

The paper is structured as follows. Section 2 describes the diarisation system itself, sections 3 and 4 describe the data and scoring metrics used in the experiments, section 5 describes the development experiments, section 6 details the performance on the RT-04 evaluation data and plans for future work and conclusions are given in sections 7 and 8.

2. SYSTEM ARCHITECTURE

The CU RT-04 diarisation system consists of three stages. The first stage segments the data with an aim of producing acoustically homogeneous segments of speech which have bandwidth and speaker labels. Gender labelling is then performed using the first pass (P1) of an ASR system to select the most likely gender for each segment in turn. The last stage performs bandwidth and gender dependent clustering to produce the final speaker labels. These stages are described in more detail in sections 2.1 ,2.2 and 2.3 respectively.

2.1. Segmentation

The segmenter, illustrated in Figure 1, is based on a system at LIMSI [5, 6] but still incorporates some of the features of the Cambridge University RT-03s segmenter [7].

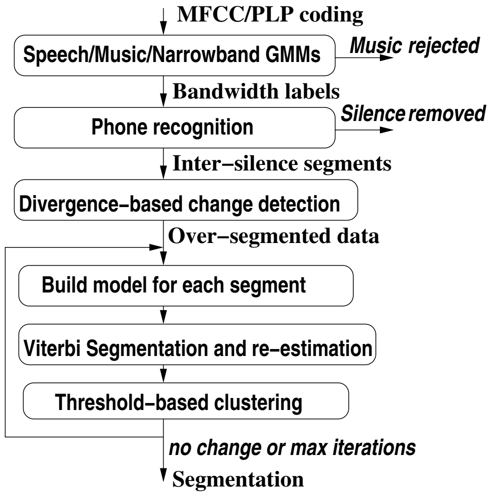

Fig. 1. The segmenter

The speech signal is coded into MFCC, wideband (WB) PLP and narrowband (NB) PLP coefficients every 10ms using a 25ms window. The data is then divided into regions of WB speech (S), speech with music (MS), NB speech (T) and music only (M) using a GMM classifier incorporating an MLLR adaptation stage, based on 13 MFCC features with first and second differentials. The MS regions are relabelled as S and the M portions are discarded. Wideband and narrowband data is subsequently treated independently.

A phone recogniser which has 45 context independent phone models per gender plus a silence model with a null language model is then run for each bandwidth. Silence portions longer than 1 second are discarded and the speech portions between these silences form the new segments. A change point detector then finds potential changes in audio characteristics within each segment. It uses a distance metric, d_{SD}, based on the symmetric Kullback Leibler (symmetric divergence) distance [8]

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

$$d_{SD} = d_{AHS} + \text{tr}\left[(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\right]$$
$$d_{AHS} = \text{tr}(\Sigma_1\Sigma_2^{-1}) + \text{tr}(\Sigma_2\Sigma_1^{-1}) - 2D$$

where D is the dimension of the feature vector, tr(x) the trace of x, μ the mean vector and Σ the covariance matrix. PLP coefficients with c0 and first differentials are used. The size of search window and the distance threshold are chosen to heavily over-segment the data ready for the subsequent phases.

These segments are then clustered into longer segments using an iterative segmentation-clustering algorithm for each bandwidth in the style of [6]. A model is built for each segment and the loss in likelihood when combining two segments is calculated from: [9]

$$\begin{aligned} d &= \frac{N_1}{2}\log|\Sigma_1| + \frac{N_2}{2}\log|\Sigma_2| - \frac{N}{2}\log|W| - \\ &\quad \frac{N}{2}\log(1 + \frac{N_1 N_2}{N^2}(\mu_1 - \mu_2)^T W^{-1}(\mu_1 - \mu_2)) \\ N &= N_1 + N_2 \\ NW &= N_1\Sigma_1 + N_2\Sigma_2 \end{aligned}$$

where Σ is the covariance matrix, μ the mean vector and N the number of frames. Segments with a loss in log likelihood less than a certain threshold are combined and then new models are built using the new segmentation which are then used to resegment the data in a Viterbi decode. This process is repeated until the segmentation does not change or a maximum number of iterations is reached. The first few iterations where there are many small segments use a single diagonal covariance model per segment, but subsequently a full covariance model is used. PLP coefficients with c0, and first and second differentials are used for this stage.

2.2. Gender Determination

Before the final clustering stage, the P1 stage of the CUHTK RT-03s ASR system [10] is used to transcribe the data. The empty segments are discarded and a forced alignment with gender dependent models is used to label the gender of each segment.

2.3. Clustering

The baseline clusterer is similar to that used in the CUED RT-03s diarisation evaluation[7] but uses the BIC-based stopping criterion introduced in [11].

The clusterer uses the start and end times of the segments from the segmenter but makes no use of the speaker labels. The clustering is done bandwidth and gender dependently using a top-down approach. Each segment is represented by a single full correlation (not covariance) matrix of 13 static PLP (with c0) features. The arithmetic harmonic sphericity distance metric[12] is used to move the segments between the children nodes until convergence before using the BIC-based stopping criterion to determine whether a given split should occur. The standard BIC formulation, given in Equation 1, is used with the slight modification that a 'local' (number of frames in the parent cluster) rather than 'global' (number of frames in the whole show) value of N is used. $\mathcal{L}$ is the log likelihood of the data, $\#M$ is the number of free parameters and α is the tuning parameter (here 7.25).

$$\text{BIC} = \mathcal{L} - \frac{1}{2}\alpha\#M\log N \qquad (1)$$

After clustering, segments with the same cluster (speaker) label which are adjacent in time are merged. This does not affect the diarisation score in itself, but makes the segmentation clearer to a reader, and enables the iterative clustering scheme of section 5.6 to be easily implemented. This *baseline clusterer* is described in more detail in [11]. The RT-04 clusterer differed only in the way the segments were sorted before clustering, changing the initialisation. Section 5.4 has more details.

3. DATA USED IN EXPERIMENTS

Four development sets were used for the experiments reported in this paper. They each consisted of roughly 30 minute extracts from 6 US news shows and are summarised in Table 1.

The didev03 set was the development data for the spring RT-03 diarisation evaluation[2] and the references were generated using the process described in [13] using forced alignments provided by the LDC with 0.3s of silence smoothing applied. The eval03 and dev04f2 sets was the official diarisation development data for the RT-04 diarisation evaluation, and were generated in a similar way to the didev03 data but used forced alignments from a LIMSI system and 0.5s silence smoothing. The sttdev04 set was marked up manually for speakers at Cambridge University and does not use the 0.5s smoothing rule, but still offers a useful development set for diarisation experiments. The key features of the development data sets are summarised in Table 1.

Name	didev03	sttdev04	eval03	dev04f2
Epoch	Oct-Dec 2000	Jan 2001	Feb 2001	Nov/Dec 2003
Spec.	RT-03s	CU	RT-04	RT-04
Alignment	LDC (words)	manual (spkrs)	LIMSI (words)	LIMSI (words)
Silence Smoothing	0.3s	N/A	0.5s	0.5s

Table 1. Summary of data sets used for development

The RT-04 diarisation evaluation data (eval04f) consisted of 12 shows broadcast in December 2003.

4. EVALUATING PERFORMANCE

4.1. Diarisation Error Rate (DER)

The diarisation error rate (DER) is the sum of the missed (speech in reference but not in hypothesis), false alarm (speech in hypothesis but not in reference) and speaker error (mapped reference and hypothesised speakers differ) rates of a system when compared to a manually defined reference. The latter is calculated by matching the hypothesised speakers to reference speakers using a one-to-one mapping which maximises the total overlap between the reference and (corresponding) hypothesis speakers. Further details can be found in [2].

A 0.25s no-score region (collar) was used round reference segment boundaries during scoring and regions of overlapping speech in the reference were excluded from scoring.

4.2. Segment Purity

The quality of the segmentation is measured by performing 'ideal' (sometimes called 'oracle') clustering on the segmenter output by assigning to each segment the true reference speaker with which it has most overlap, before scoring in the usual way. This *segment impurity* gives a measure of the miss, false alarm and within-segment speaker error, and indicates the diarisation potential from the segmentation. The number of segments must also be considered since it is possible to monotonically improve the segment purity (lower the segment impurity) by continuously splitting segments into ever smaller regions.

5. DEVELOPMENT EXPERIMENTS

5.1. Changing the Segmentation Algorithm

The segmentation described in section 2.1 was introduced into the Cambridge University diarisation system for the RT-04 evaluation. It is based on a system from LIMSI [5, 6] which was initially used in their ASR system but recently has been employed extremely successfully in their diarisation system. [14, 15]

Unlike the segmentation used in the Cambridge University RT-03 spring (RT-03s) diarisation evaluation ([7]) it produces putative speaker labels as well as the start and end times and hypothesised bandwidth of each segment. This enables a DER to be obtained after the segmentation stage. However, since there is a subsequent clustering stage in the diarisation system (which makes no use of the putative speaker labels), the most important property of the segmenter output is the segment impurity as described in section 4.2.

Results for the Cambridge University RT-03s and RT-04 segmentations are given in Table 2. They show that the change of segmenter results in a decrease in DER from 23.2% to 20.3% over the 24 development shows when using the baseline clusterer described in section 2.3.

Segment-ation	Dataset	Segment-Purity MS/FA/SPE/SI @ NumSeg	Seg DER	+Clust DER
RT-03s	didev03	0.1/3.0/1.9/5.07 @ 875	-	18.8
	eval03	0.3/1.9/1.7/3.92 @ 869	-	19.8
	sttdev04	1.0/0.9/2.1/4.01 @ 913	-	22.9
	dev04f2	1.3/4.1/1.0/6.33 @ 1077	-	32.7
	ALL	0.69/2.34/1.70/4.74 @ 3734	-	**23.2**
RT-04	didev03	0.6/1.6/1.0/3.16 @ 790	27.9	18.0
	eval03	0.6/0.7/0.9/2.17 @ 706	31.2	15.9
	sttdev04	2.2/0.3/0.9/3.36 @ 786	30.1	21.2
	dev04f2	1.5/1.8/0.6/3.93 @ 632	39.9	26.9
	ALL	1.26/1.03/0.85/3.14 @ 2914	29.7	**20.3**

Table 2. Effect of changing from the RT-03s to the RT-04 segmentation system. The % miss (MS), false alarm (FA), speaker error (SPE) and segment impurity (SI) are given, along with the number of segments after the gender-labelling phase. Also provided are the DER from the segmentation itself (where applicable) and when applying the baseline clusterer.

5.2. Changing the Likelihood Threshold in Segmentation

The final full-covariance re-segmentation stage of the segmenter uses a threshold on the log likelihood to determine which segments should be associated with the same speaker labels. The value of this threshold is critical in determining the segmenter output - too low and the data will be *oversegmented* in that too many segments will be output, whereas too high and the data will have a low segment purity as some segments will contain multiple reference speakers. The effect of changing the likelihood threshold in the full-covariance resegmentation stage is summarised in Table 3 and illustrated in Figure 2.

Lhood Thr.	Dataset	Segment-Purity MS/FA/SPE/SI @ NumSeg	Seg DER	+ base Clust	+RT04 Clust
3000	didev03	0.6/1.6/1.0/3.16 @ 790	27.9	18.0	14.0
	eval03	0.6/0.7/0.9/2.17 @ 706	31.2	15.9	15.2
	sttdev04	2.2/0.3/0.9/3.36 @ 786	30.1	21.2	22.2
	dev04f2	1.5/1.8/0.6/3.93 @ 632	39.9	26.9	23.5
	ALL	1.26/1.03/0.85/3.14 @ 2914	29.67	20.34	18.71
11000	didev03	0.6/1.6/2.6/4.82 @ 619	17.2	15.6	17.5
	eval03	0.6/0.8/1.4/2.68 @ 586	17.8	17.7	17.7
	sttdev04	2.1/0.3/2.1/4.46 @ 643	21.5	22.7	19.8
	dev04f2	1.5/1.9/1.1/4.47 @ 484	20.4	23.7	23.3
	ALL	1.23/1.06/1.82/4.10 @ 2332	**19.31**	19.95	19.45
16000	didev03	0.6/1.6/4.1/6.29 @ 578	22.7	18.9	16.1
	eval03	0.6/0.8/2.9/4.22 @ 559	21.9	16.4	17.2
	sttdev04	2.1/0.3/3.6/6.00 @ 605	24.5	20.5	20.5
	dev04f2	1.5/1.9/1.6/4.98 @ 467	15.9	**13.0**	20.0
	ALL	1.23/1.06/3.12/5.40 @ 2209	21.55	17.47	18.78
17000	didev03	0.6/1.6/4.3/6.56 @ 570	24.1	17.5	17.0
	eval03	0.6/0.8/2.6/3.96 @ 563	22.8	15.5	16.6
	sttdev04	2.1/0.3/3.7/6.10 @ 604	25.1	19.9	21.4
	dev04f2	1.5/1.9/1.8/5.15 @ 463	16.6	14.8	20.7
	ALL	1.23/1.06/3.18/5.47 @ 2200	22.46	**17.11**	18.97

Table 3. Effect of changing the likelihood threshold used in combining segments in the segmentation stage. The % miss (MS), false alarm (FA), speaker error (SPE) and segment impurity (SI) are given along with the number of segments after the gender-labelling phase. Also provided are the DER from the segmentation itself and when applying the baseline and RT-04 clusterers.

The results show that as the threshold is increased, the segment purity worsens as the number of segments decreases. The best segmenter DER is 19.31% using a threshold of 11000, with the DER of applying the baseline and RT-04 clusterers being 19.95% and 19.45% respectively. (The equivalent numbers for using static-only coefficients in the full-covariance stage are 19.15%, 21.57% and 21.21% respectively with a threshold of 2600.) The best overall performance was 17.11% for a threshold of 17000 using the baseline clusterer, the RT-04 clusterer giving 18.97% for this case. The best performance on the dev04f2 subset was 12.95% using the baseline clusterer and a threshold of 16000.

When developing the evaluation system, since the segmenter was being used as an initial stage before applying an independent clusterer, it was felt that the segmenter should try to minimise the segment impurity and hence oversegment the data. This would allow a potentially better score if improvements could be made in the subsequent clustering. For this reason a threshold of 3000 was used in the evaluation system, which led to a DER of 20.3% with the baseline clusterer and 18.7% with the RT-04 clusterer.

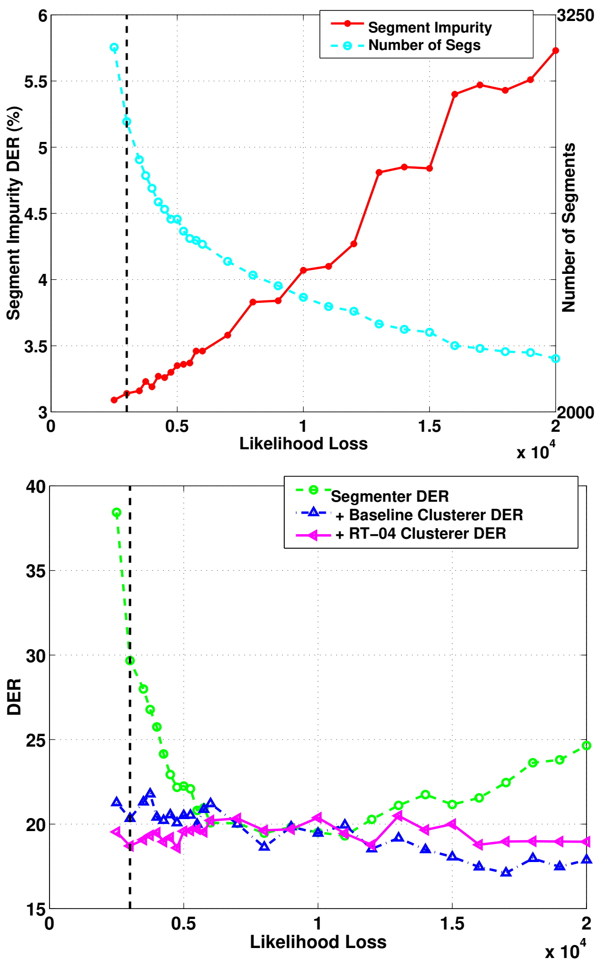

Fig. 2. Effect of changing the likelihood threshold in the final stage of the segmenter. Results show the segment impurity and number of segments, the DER of the segmenter output and the DER of the baseline and RT-04 clusterers.

5.3. Silence Removal

Silence is removed in two different places in the diarisation system. Firstly, regions of greater than a critical length which are not labelled as speech by the dual-phone recogniser are removed. This threshold is set by looking at the effect of the sum of the missed speech and false alarm speech, since these components are weighted equally in the DER. A traditional ASR system would try to have a very low miss rate, but the diarisation segmentation trades this off by allowing it to increase if the false alarm rate reduces by a greater amount.

Results of varying the silence stripping threshold are given in Table 4. The value of 1s was used as the silence threshold since this gave the lowest sum of missed and false alarm speech, and the lowest segment impurity. It also gave the lowest segmenter DER.

Empty segments after the P1 stage of the ASR system are also discarded before the final clustering stage. The effect on the miss, false alarm and segment impurity rates is given in Table 5. The number of segments over the 4 datasets is reduced by 3% with no effect on segment purity.

Silence Threshold	Dataset	Segment-Purity MS/FA/SPE/SI	@ NumSeg	Segmenter DER
0.5s	didev03	1.8/0.8/1.0/3.57	@ 1348	28.9
	eval03	2.0/0.2/0.9/3.05	@ 1229	34.8
	sttdev04	6.8/0.1/0.9/7.83	@ 1359	34.8
	dev04f2	3.3/0.5/0.6/4.37	@ 1254	47.5
	ALL	3.62/0.39/0.85/4.86	@ 5190	36.1
1s	didev03	0.6/1.6/1.0/3.21	@ 814	28.0
	eval03	0.6/0.8/0.9/2.21	@ 735	31.3
	sttdev04	2.1/0.3/0.9/3.27	@ 814	30.0
	dev04f2	1.5/1.9/0.6/3.99	@ 642	40.2
	ALL	1.22/1.08/0.85/**3.15**	@ 3005	**32.0**
2s	didev03	0.2/2.6/1.1/3.93	@ 813	29.8
	eval03	0.4/1.8/1.0/3.14	@ 770	32.4
	sttdev04	1.1/0.8/1.0/2.94	@ 804	31.9
	dev04f2	1.3/3.8/0.7/5.73	@ 658	38.2
	ALL	0.77/2.12/0.94/3.83	@ 3045	32.9

Table 4. Effect of changing the silence stripping threshold in the segmenter. The % miss (MS), false alarm (FA), speaker error (SPE) and segment impurity (SI) are given along with the number of segments *before* the gender-labelling phase.

Stage	Dataset	Segment-Purity MS/FA/SPE/SI	@ NumSeg
before P1 ASR	didev03	0.6/1.6/1.0/3.21	@ 814
	eval03	0.6/0.8/0.9/2.21	@ 735
	sttdev04	2.1/0.3/0.9/3.27	@ 814
	dev04f2	1.5/1.9/0.6/3.99	@ 642
	ALL	1.22/1.08/0.85/3.15	@ 3005
after P1 ASR	didev03	0.6/1.6/1.0/3.16	@ 790
	eval03	0.6/0.7/0.9/2.17	@ 706
	sttdev04	2.2/0.3/0.9/3.36	@ 786
	dev04f2	1.5/1.8/0.6/3.93	@ 632
	ALL	1.26/1.03/0.85/3.14	@ 2914

Table 5. Effect of removing empty segments after P1 of the ASR system. The % miss (MS), false alarm (FA), speaker error (SPE) and segment impurity (SI) are given @ the number of segments.

5.4. Initialising the Clusterer and Bandwidth Dependency

The clusterer initially assigns the segments to the children nodes based on the order they are presented. Therefore changing the order of the segments to the clusterer alters the initialisation and thus can affect the clustering results. The RT-04 segmenter assigned the speaker labels to the groups of segments somewhat arbitrarily, and initially no sorting of the segments was performed before the clustering stage. It was felt that presenting the segments in an order which kept those assigned the same cluster in the segmenter together, would be beneficial.

An experiment was therefore carried out into ways of sorting the segments before clustering. Two methods of allocating the cluster labels to the groups of segments from the segmenter were made. The first assigned the cluster labels (bandwidth and gender dependently) in ascending order using the first time of each cluster to decide the ordering. The second was similar but used the mid-time of each cluster to determine the ordering. The segments were then sorted by this new cluster-id (and by start time in the case of ties) before clustering - thus ensuring that segments assigned the same cluster-id in the segmenter would be more likely to be initialised together in the clustering stage. Contrast runs with no sorting or with purely time-based sorting were also run. The results are given in Table 6.

sorting	didev03	eval03	sttdev04	dev04f2	ALL
none	18.0	15.9	21.2	26.9	20.4
time	17.5	16.7	21.5	25.7	20.2
spkr-start	17.5	17.9	22.6	17.5	19.0
spkr-mid	14.0	15.2	22.2	23.5	**18.7**
	bandwidth dependent clustering				
none	18.3	18.6	22.4	26.9	21.4
time	18.5	15.8	20.6	25.7	20.0
spkr-start	19.4	17.9	21.3	20.0	19.7
spkr-mid	16.7	16.2	23.5	23.5	20.0
	bandwidth independent clustering				

Table 6. Effect on DER of sorting the segments before clustering. Results are presented for both bandwidth dependent and bandwidth independent clustering

Although the improvements are not consistent across the datasets, the average DER across all 24 development shows is reduced from 20.4% to 18.7% by sorting the segments by the re-assigned segmenter cluster-id and then time, before clustering. This was used for all further experiments. It is a little disturbing to note some of the variation in DER from making these changes to the initialisation. The `dev04f2` data set in particular changes from 17.5 to 23.5% just by re-allocating the initial cluster-id from its midpoint instead of its first occurrence in the show.

Table 6 also gives results for bandwidth independent clustering. This performed worse than the bandwidth dependent case, showing that automatically detected bandwidth information can be useful in distinguishing speakers.

5.5. Changing the Feature Vector

An experiment was conducted to see the effect of changing the feature vector used in the clustering stage. The Cambridge University diarisation system has always used PLP coefficients (including the cepstral c0 coefficient) but other sites have used MFCC coefficients[11, 15, 16, 17] which can sometimes perform better for diarisation[18]. The effect of changing the energy coding by using no energy coefficient, the cepstral c0 coefficient (c0), the log energy (E) and performing cepstral mean subtraction (Z) was also investigated. The results are given in Table 7. Different values of the α parameter in the stopping criterion were also tried for the different codings, but 7.25 remained the optimal in almost all cases.

The results show that performing cepstral mean subtraction considerably degrades performance, showing that the mean information *is* helping distinguish speakers. However adding both c0 and the log energy did not help improve performance. The best coding with MFCCs included the log energy but this did not perform as well as the PLP coding. The best performance overall was obtained with PLP and c0 (the standard set up) but removing the c0 coefficient improved performance on the dev04f2 data by almost 5% absolute. Further investigation showed that the shows which gained most from removing the c0 coefficient often seemed to have a low mean value for the c0 coefficient over the show. Therefore an investigation was made to see if there was a feature of the c0

Coefficients BASE	c0	E	Z	didev03	eval03	sttdev04	dev04f2	ALL
PLP	-	-	-	20.3	17.1	22.5	**18.7**	19.8
PLP	Y	-	-	14.0	15.2	22.2	23.5	**18.7**
PLP	-	Y	-	15.3	17.0	22.1	21.3	19.0
PLP	Y	Y	-	18.0	16.8	23.3	22.4	20.2
PLP	Y	-	Y	25.4	19.3	27.9	24.1	24.3
MFCC	Y	-	-	17.9	18.6	22.1	27.2	21.3
MFCC	-	Y	-	16.2	15.8	21.5	27.0	20.0
MFCC	Y	Y	-	19.7	19.3	23.5	27.4	22.4
MFCC	-	Y	Y	23.3	16.7	28.9	22.4	23.1

Table 7. Effect of changing the feature vector in clustering. Both PLP and MFCC coding were tried with combinations of c0, log energy (E) and cepstral mean subtraction (Z).

coding which might help predict whether the c0 coefficient should be used in clustering for optimal performance.

5.5.1. *c0 switching*

It had been observed that usually including c0 in the feature vector improved clustering, but some times it did not. Experiments were performed to see if a property of the c0 coefficient itself could be used to predict whether this gain would occur. Five properties of the c0 coefficient were investigated, namely the mean value of the data for the show after segmentation *(mean(show))*, the mean value of the segment means *(mean(segmean))*, the standard deviation of the segment means *(stddev(segmean))* and the ratios of the latter two. The correlation coefficients between the showwise difference in DER from including c0 and the property in question is given in Table 8.

Property	Correlation
stddev(segmean)	-0.0295
mean(segmean)/stddev(segmean)	0.0995
stddev(segmean)/mean(segmean)	-0.2223
mean(segmean)	0.4223
mean(show)	0.4560

Table 8. Correlation Coefficients between the c0 property and the difference in DER from including c0 in the clustering for all 24 development shows

The correlation coefficients show that the most correlated feature is the mean value of the c0 coefficient across the whole show after segmentation, with a correlation of 0.456. Figure 3 shows a scatter plot of the mean c0 value against the difference in DER when including c0 and the mean DER across all 24 development shows when the clustering uses c0 if and only if the mean c0 value after segmentation is above a certain threshold. The breakdown in results over the different datasets is given in Table 9.

The results show the mean DER over the 24 shows can be reduced from 18.7% to 17.7%, with the DER on the dev04f2 dataset (closest in epoch to the eval04f data) reduced from 23.5% to 19.2% if this method is used with a threshold of 50. However, there was some concern that this may not hold across new datasets, so the c0-switching was implemented as a contrast run for the RT-04 evaluation, the primary run using c0 in the clustering stage for all cases.

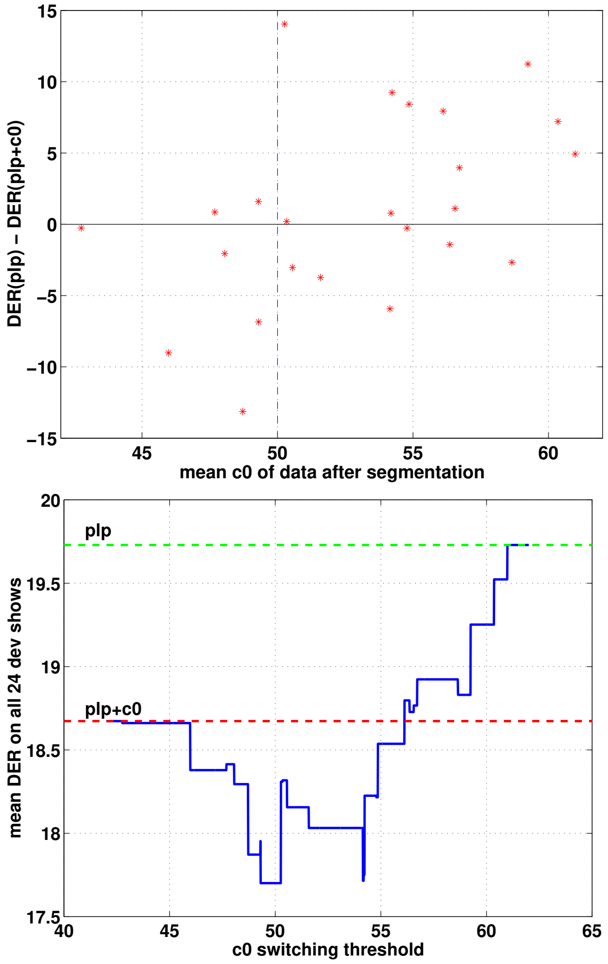

Fig. 3. (a) Scatter plot showing the difference in DER when omitting the c0 coefficient against the mean c0 value for each development show. (b) Mean DER across all 24 development shows when only including c0 in clustering if the mean value is above a critical threshold.

c0thresh	didev03	eval03	sttdev04	dev04f2	ALL
0 (PLP+c0)	14.0	15.2	22.2	23.5	18.7
48	14.0	15.2	22.2	22.3	18.5
49	14.0	15.2	21.8	20.3	17.9
50	14.0	15.5	21.8	19.2	**17.7**
51	16.6	15.5	21.2	19.2	18.2
52	16.6	15.5	21.2	18.7	18.1
54	16.6	15.5	21.2	18.7	18.1
56	18.6	15.5	21.2	18.7	18.6
100 (PLP)	20.3	17.1	22.5	18.7	19.8

Table 9. Results per dataset from only including c0 in the clustering if the mean value of the show is greater than a threshold

5.5.2. *Using Delta Features*

An experiment was conducted which added first differentials (deltas) to the feature vector but used a block diagonal covariance representation in the clustering. The results for different α values on the development datasets are illustrated in Figure 4. The optimal α value is much lower here than for the static only case (as it is influenced by the independence of the features), and the best performance is only 20.6% compared with the 18.7% from the static-only case, therefore this was not used in the evaluation system.

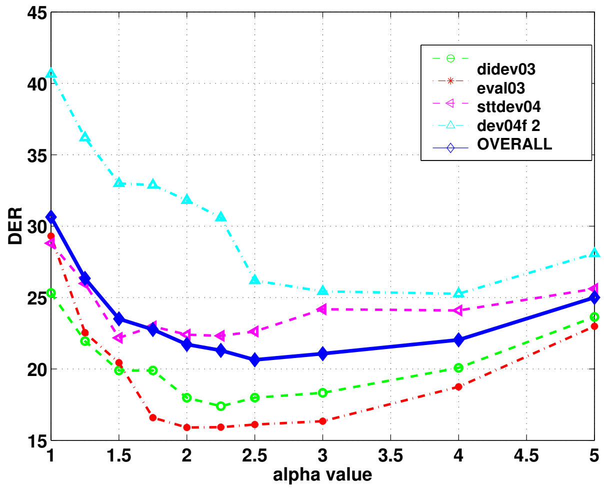

Fig. 4. Effect of changing the α value in the clustering when using a block diagonal representation with static and delta coefficients.

5.6. Iterative Clustering

Iterative clustering or re-segmentation could potentially help improve the performance of diarisation systems. We implemented a simple iterative scheme which ran the clusterer and then merged temporally adjacent segments which were clustered together, before running the clusterer again on the new segmentation. The idea was that segments which are adjacent in time are often spoken by the same speaker and thus if the clusterer also clustered them together then there are two sources suggesting the segmentation should be refined to combine the segments in question.

The final clustering stage is run as before, but the preceding clustering stages can be run differently if required. For example, producing many clusters would minimise the risk of segments being falsely combined, whereas producing fewer clusters than normal and relying on the temporal adjacency criterion to restrict false combinations might also be justified.

The results for α of 7.25 (optimal), 5 (conservative) and 10 (overclustered) for the non-final iteration are presented in Table 10 and the segment purity for the case of using the optimal $\alpha = 7.25$ throughout is given in Table 11.

The results show that this technique did not help improve performance overall, producing a increase in segment impurity and corresponding increase in final DER even after only one extra iteration, for all datasets except the `eval03` data.

5.7. Varying the Parameters

Finally, the α value and the decision to use the 'local' or 'global' formulation of the BIC stopping criterion [11] was checked. The results, illustrated in Figure 5, confirm that the best result of 18.7% occurs using α=7.25 with the 'local' formulation.

non-final α	itera-tions	eval03	didev03	sttdev04	dev04f2	OVERALL
-	0	63.3	59.7	62.4	67.5	63.1 @ 2914
-	1	15.2	14.0	22.2	23.5	18.7 @ 2629
7.25	2	14.9	15.9	22.6	23.6	19.3 @ 2587
7.25	3	15.6	14.8	22.3	23.7	19.1 @ 2570
7.25	10	15.6	15.0	22.0	23.7	19.1 @ 2565
5	2	15.3	15.9	23.1	24.3	19.7 @ 2609
5	3	16.7	17.3	24.3	28.9	21.7 @ 2616
5	10	16.4	18.6	23.6	28.1	21.6 @ 2621
10	2	15.0	17.5	21.6	23.1	19.3 @ 2540
10	3	16.9	17.8	24.6	24.0	20.9 @ 2521
10	10	19.9	21.3	25.3	22.9	22.4 @ 2515

Table 10. Iterative clustering merging temporally adjacent segments in the same cluster between stages. Results show the final DER @ the number of segments.

iter	didev03	eval03	sttdev04	dev04f2	ALL
0	1.0 @ 790	0.9 @ 706	0.9 @ 786	0.6 @ 632	0.85 @ 2914
1	1.6 @ 714	0.9 @ 644	1.9 @ 702	1.2 @ 569	1.43 @ 2629
2	2.1 @ 694	1.0 @ 638	2.0 @ 695	1.2 @ 560	1.59 @ 2587
3	2.2 @ 686	1.4 @ 635	2.2 @ 691	1.2 @ 558	1.80 @ 2570
10	2.3 @ 681	1.4 @ 635	2.2 @ 691	1.2 @ 558	1.82 @ 2565

Table 11. Segment impurity *excluding the MS and FA components* @ number of segments for the iterative clustering with $\alpha = 7.25$.

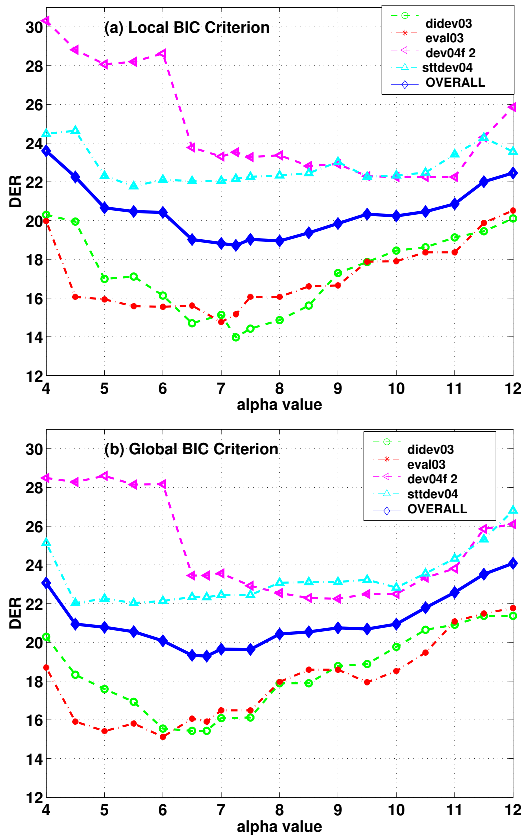

Fig. 5. Effect of changing the α value in the clustering when using the full correlation matrix with static only PLP coefficients. (a) uses the 'local' whilst (b) uses the 'global' formulation.

6. RESULTS ON THE RT-04 EVALUATION DATA

Table 12 shows the results on the 12-show RT-04 evaluation data (`eval04f`) and the progress in diarisation at Cambridge University since the RT-03s evaluation. Introducing the new clustering[11] reduced the primary DER from 36.3% to 27.9%, whilst subsequently introducing the new segmenter reduced this further to 22.5%. It was discovered after the evaluation that the coding into PLP coefficients had been affected by switching compilers despite no change to the source code, and this had unfortunately led to an increase of DER to 23.9%. This confirms the observation in section 5.4 that the clustering is somewhat over-sensitive to slight changes in input, possibly due to the system being top-down instead of using the more common agglomerative method.

Coding	Segmentation	Clustering	DER main	DER c0switch
RT-03s	RT-03s	RT-03s	36.33	-
RT-03s	RT-03s	RT-04	27.90	24.45
RT-03s	RT-04	RT-04	22.48	22.35
† RT-04	RT-04	RT-04	23.86	24.12

Table 12. Progress since RT-03s on the `eval04f` data. The DER of the main system is given along with the contrast run with the c0-switching where applicable. † Official eval. submission (see [4]).

The contrast run which included the c0-switching did perform better when using the RT-03s segmentation (24.5% instead of 27.9%) but made little difference when used with the RT-04 segmenter.

An experiment was run to see the effect of using different criteria to pick the likelihood threshold and clustering strategy on the dev data. Three different strategies were tried namely (a) just use the segmenter output which gave the best segmenter DER; (b) use the clusterer output which gave best performance across all the dev data; and (c) use the clusterer output which gave best performance on the `dev04f2` data, since this was closest in epoch to the `eval04f` data. The results are given in Table 13 using the RT-03s PLP coding.

likeli-hood thresh	post-ASR Segment Impurity	Segmenter DER	Baseline Clusterer DER	RT-04 Clusterer DER
3000	0.4/1.1/1.2/2.69 @ 1383	35.15	22.03	22.48(e)
11000	0.4/1.1/2.2/3.73 @ 1063	18.72(a1)	22.90	21.02
16000	0.4/1.1/3.8/5.35 @ 987	21.17	20.50(c)	22.18
17000	0.4/1.1/3.9/5.44 @ 988	22.05	22.06(b)	21.44
†2600	0.4/1.1/3.5/5.05 @ 979	18.12(a2)	21.82	23.49

Table 13. Effect on `eval04f` DER of using different criteria on the dev data to choose the segmenter likelihood threshold and clustering strategy. (e) represents the RT-04 evaluation system, (a1) and (a2) are from the optimal segmenter DER on the dev data, (b) the optimal clustered DER on the dev data, and (c) the optimal clustered DER on the `dev04f2` data. † no differentials used in the feature vector in the full-covariance stage of the segmenter.

The results show that the `eval04f` DER could have been reduced by changing the strategy used to finalise the system on the dev data, the best performance being 18.1% when using the segmenter output directly (with no differentials in the feature vector in the full-covariance stage).

7. FUTURE WORK

Future work will look at trying to use multiple knowledge sources to improve the diarisation system, for example by using the speaker labels from the segmenter within the clusterer, or combining segmenter and clusterer outputs using cluster voting[19, 20]. More information from the ASR system may also be incorporated, as in [21]. The use of proxy speaker models[22] which has been successfully implemented within the diarisation framework at MIT[23] will also be investigated, along with the use of 'standard' speaker identification techniques, which give large benefits in the LIMSI RT-04 diarisation system[15]

8. CONCLUSIONS

This paper has described the Cambridge University RT-04 diarisation system, including details of the new segmentation and clustering components. Many experiments made to try to improve the performance of the system have been reported although few affected the final system. The clustering component was rather sensitive to the segmentation, with small changes in input often making large changes in results. The final system gave a diarisation error rate of 23.9% on the RT-04 evaluation data, a 34% relative improvement over the Cambridge University RT-03s system, and it was shown that this score could have been reduced further to 18.1% within this diarisation framework.

9. ACKNOWLEDGEMENTS

The authors would like to thank Kit Thambiratnam for work in broadcast news segmentation whilst at Cambridge University.

10. REFERENCES

[1] NIST, "Benchmark Tests : Rich Transcription (RT)," http://www.nist.gov/speech/tests/rt/.

[2] NIST, "The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan, version 4," http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf, 25th February 2003.

[3] NIST, "Fall 2004 Rich Transcription (RT-04F) Evaluation Plan," http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf, 30th August 2004.

[4] J. G. Fiscus, J. S. Garofolo, A. Le, A. F. Martin, D. S. Pallett, M. A. Przybocki, and G. Sanders, "Results of the Fall 2004 STT and MDE Evaluation," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04F)*, November 2004, p. to appear.

[5] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and Transcription of Broadcast News Data," in *Proc. ICSLP*, December 1998, vol. 4, pp. 1335–1338.

[6] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, May 2002.

[7] S. E. Tranter, K. Yu, D. A. Reynolds, G. Evermann, D .Y. Kim, and P. C. Woodland, "An Investigation into the Interactions between Speaker Diarisation Systems and Automatic Speech Transcription," Tech. Rep. CUED/F-INFENG/TR-464, Cambridge University Engineering Dept., Oct. 2003.

[8] P. J. Moreno and P. P. Ho, "A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels," Tech. Rep. HPL-2004-7, HP Laboratories Cambridge, January 9th 2004.

[9] H. Gish, M.-H. Siu, and R. Rohlick, "Segregation of Speakers for Speech Recognition and Speaker Identification," in *Proc. ICASSP*, April 1991, vol. 2, pp. 873–876.

[10] D. Y. Kim, G. Evermann, T. Hain, D. Mrva, S. E. Tranter, L. Wang, and P. C. Woodland, "Recent Advances in Broadcast News Transcription," in *Proc. ASRU*, December 2003, pp. 105–110.

[11] S. E. Tranter and D. A. Reynolds, "Speaker Diarisation for Broadcast News," in *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2004, pp. 337–344.

[12] F. Bimbot and L. Mathan, "Text-Free Speaker Recognition using an Arithmetic Harmonic Sphericity Measure," in *Proc. Eurospeech*, September 1993, vol. 1, pp. 169–172.

[13] NIST, "Reference Cookbook for "Who Spoke When" Diarization Task, v2.4," http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/ref-cookbook-v2_4.pdf, 17th March 2003.

[14] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, C. Barras, L. Chen, F. Lefevre, S. Meignier, and A. Messauodi, "Summary of Progress at LIMSI," in *EARS Mid-Year Meeting*, February 2004.

[15] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving Speaker Diarization," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04F)*, November 2004, p. to appear.

[16] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The ELISA Consortium Approaches in Broadcast News Speaker Segmentation during the NIST 2003 Rich Transcription Evaluation," in *Proc. ICASSP*, May 2004, vol. 1, pp. 373–376.

[17] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards Robust Speaker Segmentation: The ICSI-SRI Fall 2004 Diarization System," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04F)*, November 2004, p. to appear.

[18] C. Wooters, "Speaker-Attributed STT. Who Spoke the Words," in *Proc. Fall 2003 Rich Transcription Workshop (RT-03f)*, November 2003.

[19] S. E. Tranter, "Cluster Voting for Speaker Diarisation," Tech. Rep. CUED/F-INFENG/TR-476, Cambridge University Engineering Department, May 2004.

[20] S. E. Tranter, "Two-way Cluster Voting to Improve Speaker Diarisation Performance," in *Proc. ICASSP*, March 2005, p. to appear.

[21] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker Diarization from Speech Transcripts," in *Proc. ICSLP*, October 2004, pp. 1272–1275.

[22] Y. Akita and T. Kawahara, "Unsupervised Speaker Indexing using Anchor Models and Automatic Transcription of Discussions," in *Proc. Eurospeech*, September 2003, vol. 4, pp. 2985–2988.

[23] P. A. Torres-Carrasquillo and D. A. Reynolds, "The MIT Lincoln Laboratory Speaker Diarization Systems," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04F)*, November 2004, p. to appear.