Speaker Diarisation for Broadcast News

S. E. Tranter † and D. A. Reynolds ‡

† Cambridge University Engineering Department
Trumpington Street, Cambridge
CB2 1PZ, UK
sej28@eng.cam.ac.uk

‡ MIT-Lincoln Laboratory
244 Wood Street, Lexington
MA 02420-9185, USA
dar@ll.mit.edu

Abstract

It is often important to be able to automatically label 'who spoke when' during some audio data. This paper describes two systems for audio segmentation developed at CUED and MIT-LL and evaluates their performance using the speaker diarisation score defined in the 2003 Rich Transcription Evaluation. A new clustering procedure and BIC-based stopping criterion for the CUED system is introduced which improves both performance and robustness to changes in segmentation. Finally a hybrid 'Plug and Play' system is built which combines different parts of the CUED and MIT-LL systems to produce a single system which outperforms both the individual systems.

1. Introduction

Segmenting audio data into speaker-labelled regions has many applications, including improving readability of transcripts, enabling speakers to be tracked, conversations to be followed, data to be indexed, browsed or searched by speaker, and aiding speaker adaptation techniques in speech recognition.

A particularly challenging domain for speaker labelling is broadcast news shows. These programs contain an unpredictable number of speakers who speak for a wide range of different times, sometimes simultaneously; as well as containing unwanted regions such as commercial (advert) breaks. However, tracking speakers through current affairs debates, or being able to search for information known to be spoken by the primary anchor or newsreader, can be very beneficial - and so the task of identifying 'who spoke when' in broadcast news audio is particularly interesting and challenging.

This paper describes systems developed at CUED and MIT-LL to perform automatic segmentation, clustering and labelling of speakers (and in some cases commercial breaks) in broadcast news data. The paper is arranged as follows: the 'diarisation' error rate used for scoring is explained in Section 2 and the data used for experiments defined in Section 3. The December 2003 CUED diarisation system is described in Section 4 which also introduces a new clustering procedure with new stopping criteria. The MIT diarisation system is described in Section 5 and a hybrid 'Plug and Play' system which combines stages of both the CUED and MIT-LL system is described in Section 6 along with comprehensive experimental results. Finally conclusions are offered in Section 7.

This work was supported by DARPA grant MDA972-02-1-0013† and the Department of Defense under Air Force contract F19628-00-C-0002 ‡. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. Diarisation

Speaker diarisation was a task in the 2003 Spring Rich Transcription (RT-03s) evaluation.[1] This was a development of the RT-02 Broadcast News Metadata speaker segmentation task [2] which in turn developed from the (multi-speaker) speaker segmentation task in the speaker recognition evaluations.[3, 4]

A system hypothesises a set of speaker segments each of which consist of a speaker-id label and the corresponding start and end time. This is then scored against a reference 'ground-truth' speaker segmentation. A one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs is performed so as to maximise the total overlap of the reference and (corresponding) mapped hypothesis speakers. Speaker detection performance is then expressed in terms of the miss (speaker in reference but not in hypothesis), false alarm (speaker in hypothesis but not in reference), and speaker-error (mapped reference speaker is not the same as the hypothesised speaker) rates. The overall diarisation score is the *sum of these three components*, and can be calculated using the following formula:

$$\frac{\sum_s dur(s) \cdot (\max(N_R(s), N_H(s)) - N_C(s))}{\sum_s dur(s) \cdot N_R(s)} \quad (1)$$

where s is the longest continuous piece of audio for which the reference and hypothesised speakers do not change, $dur(s)$ is the duration of s, $N_R(s)$ is the number of reference speakers in s, $N_H(s)$ is the number of hypothesised speakers in s and $N_C(s)$ is the number of mapped reference speakers which match the hypothesised speakers. Since the RT-03s diarisation score excluded from scoring areas where multiple reference speakers were talking simultaneously, and we do not postulate any regions of overlapping speech in the hypotheses, this formula becomes:

$$\frac{\sum_s dur(s) \cdot (H_{miss}(s) + H_{fa}(s) + H_{spe}(s))}{\sum_s dur(s) \cdot H_{ref}(s)} \quad (2)$$

where H is always zero except $H_{miss}(s)$ is 1 for a missed speech segment, $H_{fa}(s)$ is 1 for a false alarm speech segment, $H_{spe}(s)$ is 1 for a segment with a speaker error, and $H_{ref}(s)$ is 1 for a segment containing a reference speaker.

The references used for scoring were generated according to the rules specified in [1, 5]. Effectively, speaker turns were derived using word times generated by a word-level forced alignment from the Linguistic Data Consortium (LDC), with segment breaks when either a new speaker starts talking, or the speaker pauses for more than a certain critical length of time (here fixed at 0.3s as was used in the RT-03s diarisation evaluation). Speaker-attributable non-lexical events, such as {cough, breath, lipsmack, sneeze and laughter} were excluded from scoring along with their adjoining silences. Commercial breaks were not transcribed for the reference, and as a result were also excluded from scoring in the primary scoring metric, although we also consider a secondary metric which penalises systems for retaining adverts in their hypothesised output.

3. Data used in experiments

The experiments reported in this paper were conducted on the diarisation development data *(bndidev03)* and the entire 2003 evaluation data *(bneval03)* used in the English Broadcast News RT-03 Rich Transcription evaluations.[6]

Each data set consists of one 30 minute extract from 6 different US broadcast news shows. Two of these are radio shows, namely Voice of America English News (VOA_ENG) and PRI The World (PRI_TWD); and four are TV shows, namely NBC Nightly News (NBC_NNW), ABC World News Tonight (ABC_WNT), MSNBC News with Brian Williams (MNB_NBW) and CNN Headline News (CNN_HDL). Details of the exact composition of the data sets can be found in [7].

4. CUED diarisation system

The CUED December 2003 diarisation system can be split into three basic components. Firstly there is an optional stage of advert detection, namely trying to postulate where commercial breaks occur within the broadcast news shows. Next the remaining data is segmented, which aims to produce acoustically homogeneous segments of speech with bandwidth and gender labels. Finally clustering is performed to group together segments from the same speaker to produce the final speaker labels. This process is illustrated in Figure 1.

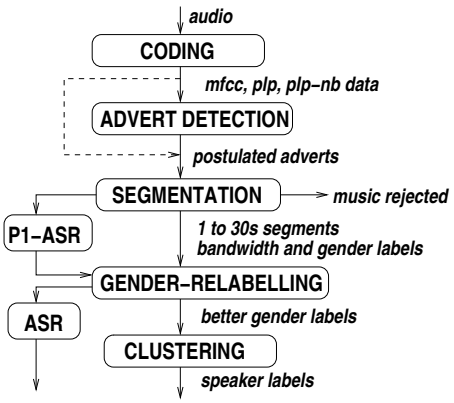

Figure 1: The CUED Dec 2003 diarisation system

4.1. Advert removal

The advert detection stage is similar to that used in the TREC-8 Cambridge Spoken Document Retrieval system[8]. It uses a direct search of the audio, as described in [9] to find exact matches which represent re-broadcast (pre-recorded) portions of the news shows. These repeats are then converted into postulated commercial breaks by applying a series of rules relating to the number of times the audio is repeated and the gaps between labelled repeats.

A library of broadcast news shows was made[1] using the English TDT-4 training data, excluding the shows from the RT-03s development sets. This consisted of between 40 and 70 shows for each of the 6 broadcasters spanning October 2000 to January 2001. This library is denoted `CU_TDT4`. A further library was generated which excluded shows for each broadcaster which were broadcast in the same calendar month as that broadcaster's episode in the diarisation development data. This was to simulate conditions in the RT-03s evaluation, where there was a temporal gap between the test-audio and the training shows. This library is denoted `CU_EVAL`.

The data for both the library and the evaluation shows is first coded at a frame rate of 100Hz into 39-dimensional feature vectors consisting of the normalised log-energy and 12 Mel-frequency PLP cepstral parameters along with their first and second derivatives.

Overlapping windows are generated on the data; 5 seconds long with a 1 second shift for the ABC, CNN, MNB and NBC shows, and 2.5 seconds long with a 0.5 second shift for the VOA and PRI shows. The difference in these values reflects the nature of the shows, the radio shows in general having fewer well-defined commercial breaks, but still including other repeated material such as station jingles which could be removed automatically. The windows are then represented by a diagonal correlation matrix. (It was found that using the correlation matrix instead of the covariance matrix gave better results due to the retention of the mean information.)

The Arithmetic Harmonic Sphericity (AHS) distance[10] is then calculated for each evaluation window compared to each library window. This is marked as a repeat if this distance metric falls below a small threshold. For a perfect match the distance would be zero, but since the granularity of the windows means there may be a delay of up to half the window shift between corresponding events in the two audio streams, causing a slight mismatch in the data, a threshold is required. This is set conservatively so that there should not be any false matches whose distance metric is lower than the threshold. To remove any false positives, and guard against the possibility of a news-story being repeated on different shows, the evaluation window had to match at least 2 different library windows to be marked as a repeat.

After finding the repeats, smoothing was carried out between the areas labelled as repeats in order to identify the commercial breaks. The smoothing relabelled any audio of less than a certain duration which occurred between two repeats

[1]The library used in the advert detection was obtained automatically by windowing over whole training shows. It is possible in theory to manually mark the training data to define a true 'library of known adverts' but this was considered impractical on these large data sets.

as part of the adverts unless this made the overall commercial break exceed a maximum duration. These values were chosen on a broadcaster-specific basis to reflect the overall properties of the broadcasts, but in general the maximum permitted duration was around 3 minutes, and the smoothing for the TV shows was just over 1 minute, with minimal smoothing for the radio shows. CNN had less smoothing than the other TV sources due to the frequent occurrence of 20 to 30s long sports reports between adverts and station jingles. Finally the boundaries of the postulated commercial breaks were refined to take into account the granularity of the initial windowing.

Further details and analysis of the effectiveness of this technique can be found in [7]. The CU_TDT4 system removed 18.4% of the audio, which consisted of 1783s=86.3% of all the adverts and 198s=2.28% of all the news; whilst the CU_EVAL system removed 6.75% of the audio, which consisted of 582s=28.2% of all the adverts and 144s=1.66% of all the news on the diarisation development data. The system removed 8.9% of the evaluation data, which consisted of 867s=40.5% of all the adverts and 70s=0.83% of all the news.

4.2. Segmentation

The CUED RT-03s segmentation was used for these experiments[7]. The data is first coded at a frame rate of 100Hz into 39-dimensional feature vectors consisting of the normalised log-energy and 12 MFCC coefficients along with their first and second derivatives. This data is then run through a GMM classifier which has models for wideband speech (S), telephone speech (T), speech with music/noise (MS) and pure music/noise (M). The MS segments are relabelled as S and the M portions discarded, leaving bandwidth labelled data. An inter-class transition penalty is used which forces the classifier to produce longer segments and an additional penalty on leaving the M model reduces the number of misclassifications of speech as music. The classification also includes an adaptation stage, using MLLR to adapt both the means and variances of the models using the first stage classification as supervision.

A phone recogniser, which has 45 context independent phone models per gender plus a silence/noise model with a null language model is then run for each bandwidth separately. The output of the phone recogniser is a sequence of phones with male, female or silence tags. The phone identifiers are ignored but the phone sequences with the same gender are merged and some heuristic smoothing rules applied to produce a series of small segments, using the silence tags to help define the boundary locations.

Finally clustering and merging of similar temporally adjacent segments is performed using the GMM classifier output to restrict the boundary locations, to produce the final segmentation with bandwidth and putative gender labels. The final gender labels are produced by aligning the output of the first-pass of the CUED RT-03 Broadcast News ASR system [11] with gender dependent models. The segments are then assigned to the gender which gives the highest likelihood.

4.3. Clustering

Segment clustering is performed on the segments separately for each bandwidth and gender, making the assumptions that the gender and location of a speaker will not change within a broadcast; and that these properties can be labelled with sufficient accuracy to aid clustering performance.

Each segment is represented by a full *correlation* matrix of the 13-dimensional PLP vectors (*without* first or second derivatives) and the distance metric used is the Arithmetic Harmonic Sphericity (AHS).[10] The clustering is performed top-down as follows:

0. Initialise all segments into a single active node

foreach active node:

1. Assign its segments to one of two children nodes, maintaining the order to exploit the temporal closeness.
2. Calculate the Gaussian statistics for the 2 children nodes.
3. Move any segment which gives a smaller distance to the sibling node than its own node.
4. Update the node statistics.
5. Repeat steps 1-4 until no segments move or the maximum number of iterations is reached.
6. If the stopping criterion is not met add the children nodes to the active node list and remove the parent node, otherwise turn the parent node into a leaf node.

The stopping criteria are critical in determining the final clusters. The system allows several different criteria to be used which reflect the aim of the clustering. These include specifying a minimum occupancy for clusters (used in the ASR system where a certain amount of data is necessary for adaptation, but not for diarisation where speakers can talk for arbitrarily short portions of time) or using measures based on the 'cost' as defined by the average distance of the segments from the nodes. For this paper we also implemented a new stopping criterion based on the Bayesian Information Criterion.

4.3.1. Cost-based stopping criterion

The clustering used in the CUED RT-03s diarisation system was based on a *4-way* (not 2-way) splitting algorithm.[7] If the parent node could not be split into 4 in a way that satisfied the stopping criteria, the child segments would be merged to try to form a 3-way split, and if this failed a 2-way split was attempted. If this was also disallowed, the parent node became a leaf node.

Three parameters were used to control the cost-based stopping criteria. The first was the most important, and specified the ratio of the gain in cost function from splitting to the global node cost. A node cost is the sum of the distances of its segments to itself and the gain in splitting is the cost of the parent node minus the cost of the children nodes. We call this the '*h*-parameter'. Additionally the '*p*-parameter' controlled the ratio of the inter:intra child cost and the '*j*-parameter' provided a multiplicative component used to weight the scores for the special case of a node containing only one segment since the distances are zero in this case. This system gave a diarisation error rate of 33.29% on the development (bndidev03) data and 32.30% on the evaluation (bneval03) data.

The CUED December 2003 system discussed in this paper uses a simpler 2-way splitting algorithm. Therefore, the '*p*-parameter' is set arbitrarily high and the '*j*-parameter' is set to 1 since they are not as important for the 2-way splitting procedure. The cost-based stopping criterion is thus controlled solely by varying the '*h*-parameter'.

The results from varying the *h*-parameter on the diarisation development (bndidev03) and evaluation (bneval03) data are illustrated in Figure 2, showing the method generalises reasonably well to the unseen evaluation data.

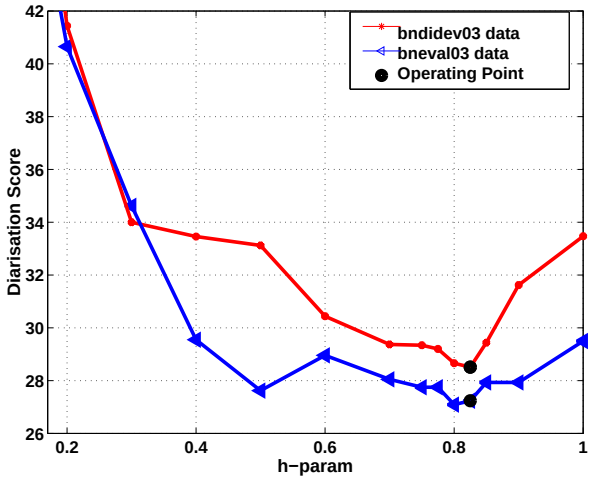

Figure 2: Effect of changing the h-parameter on the diarisation score on the development (bndidev03) and evaluation (bneval03) data.

4.3.2. *BIC-based stopping criteria*

An alternative stopping criterion, commonly used in speaker clustering, is based on the Bayesian Information Criterion (BIC)[12, 13]. This is effectively a penalised log likelihood:

$$\text{BIC} = \mathcal{L} - \frac{1}{2}\alpha \# M \log N \tag{3}$$

where $\#M$ is the number of free parameters, N the number of data points and α the tuning parameter, usually set to 1. The data is modelled using a full Gaussian of dimension d:

$$p(x_k) = \frac{1}{(2\pi)^{0.5d}|S|^{0.5}} \exp^{-0.5(x_k-\mu)'S^{-1}(x_k-\mu)}$$

where μ is the mean vector, S is the covariance matrix and $|S|$ is the determinant of S . The log likelihood term, $\mathcal{L}$ is then

$$\mathcal{L}(x_1,..,x_N) = \sum_{k=1}^{N} \log p(x_k) = -\frac{1}{2} N \log(|S|) + NC$$

where C is a constant, $-\frac{1}{2}d(1+\log(2\pi))$. The number of free parameters for K clusters is:

$$\#M = K\left(d + \frac{d(d+1)}{2}\right)$$

Thus when making a local decision as to whether a cluster Z should be split into 2 clusters, X and Y, the equations become:

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

$$\text{BIC}_{x+y} = -\frac{1}{2}[N_x \log(|S_x|) + N_y \log(|S_y|)] - 2\alpha P + N_z C$$

$$\text{BIC}_z = -\frac{1}{2} N_z \log(|S_z|) - \alpha P + N_z C$$

$$\Delta\text{BIC} = \frac{1}{2}[N_z \log(|S_z|) - N_x \log(|S_x|) - N_y \log(|S_y|)] - \alpha P$$

and the split goes ahead if $\Delta\text{BIC} > 0$. We call this formulation *BIC-local* as the decision about whether to split a particular cluster is taken locally. Alternatively the whole cluster set can be viewed as an entity, and the decision then becomes should the K clusters be increased to $K+1$. In this case the formula for the ΔBIC remains the same *except* that the N used in the penalty term, P, becomes the total number of frames in all the clusters, N_f, rather than the number in the cluster being split, N_z. We call this formulation *BIC-global*.

In general the BIC formula is used in conjunction with agglomerative clustering, so can be thought of as a decision as to whether to merge the two clusters X and Y into Z (rather than splitting Z into X and Y). In this case, the choice of which clusters to merge is usually made such as to produce the most negative ΔBIC. If this is non-negative the merge does *not* go ahead and all clustering is stopped.

The CUED implementation instead uses a divisive clustering scheme which tries to split each active node in turn and does not order the decisions. The segment assignment for a given potential split is made as before using the full correlation matrix and the AHS distance, but the decision as to whether to split a node is now taken by testing if the ΔBIC is > 0. For this reason it was felt that the BIC-local formulation may be more appropriate for this case.

The results from changing the α penalty using both the BIC-global and BIC-local implementations on the development (bndidev03) and evaluation (bneval03) data are illustrated in Figure 3. The BIC-global implementation seems to generalise slightly better, with the performance across both data sets roughly matching each other except for one point and the same value of α producing the best performance in both cases. However, the BIC-local implementation, although slightly more noisy, does give slightly better performance.

4.3.3. *Summary*

A summary of the performance on the development (bndidev03) and evaluation (bneval03) data for the optimal parameter values of the three different stopping criterion on the two data sets is given in Table 1. The results show that the best performance is produced using the BIC-local implementation, but if the parameters are tuned on the development data there is little difference in performance between the two BIC implementations on the evaluation data.

The new 2-way clustering strategy with the introduction of the BIC stopping criteria has reduced the diarisation error by $7 \rightarrow 8\%$ absolute compared to the CUED RT-03s evaluation system[7] on both the development and evaluation data.

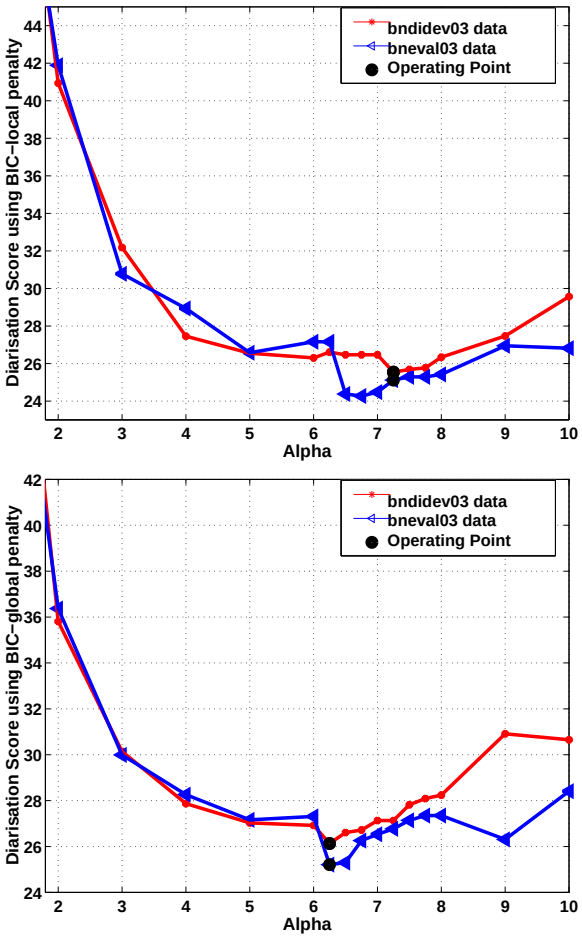

Figure 3: Effect of changing α on the diarisation score on the development (bndidev03) and evaluation (bneval03) data. The first graph is using BIC-local and the second BIC-global

Stopping Criterion	Optimal Param		Diarisation Score	
	bndidev03	bneval03	bndidev03	bneval03
RT-03s sys	-	-	33.29	32.30
Cost-based	0.825	-	28.51	27.24
Cost-based	-	0.8	28.66	27.09
BIC-global	6.25	6.25	26.13	25.21
BIC-local	7.25	-	25.54	25.12
BIC-local	-	6.75	26.47	24.27

Table 1: Optimal parameters and performance using the different stopping criterion.

5. MIT-LL diarisation system

The MIT-LL RT-03s BN diarisation system, shown in Figure 4, consists of three main components: An initial segmentation to detect putative change points in the audio stream, a classification of these segments as speech or non-speech, and a clustering stage to associate speech segments with each speaker present in the audio file. In addition to the main components, there is also a speech activity detection (SAD) gating stage and a gender classification on the final segmentations. The system is described in more detail in [7].

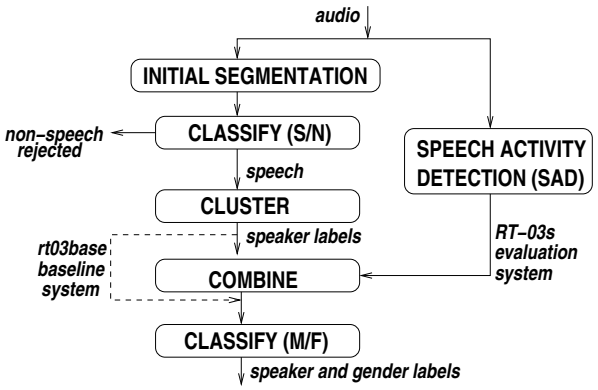

Figure 4: The MIT-LL BN diarisation system

This system gave a diarisation error of 24.46% on the development and 23.85% on the evaluation data.

5.1. Segmentation

The initial segmentation is based upon a BIC change point detection algorithm[13]. The audio signal is first converted into a stream of feature vectors at a frame rate of 100Hz consisting of 30 MFCC coefficients extracted over the full 8kHz bandwidth. No channel compensation is applied so as to exploit differences in channels to aid in detection of change points in the audio signal. For a window of N feature vectors, $\{x_1, x_2, ..., x_i, ..., x_N\}$, the BIC statistic is computed for all possible change points i in the window :

$$\Delta\text{BIC}(i) = -\log \frac{p(X/\lambda)}{p(X_1/\lambda_1)p(X_2/\lambda_2)} - \alpha P \qquad (4)$$

where $X_1 = \{x_1, ..., x_i\}$, $X_2 = \{x_{i+1}, ..., x_N\}$, $X = \{x_1, ..., x_N\}$, and $\lambda, \lambda_1, \lambda_2$ are full covariance Gaussian models trained with X, X_1, X_2 respectively. P and α are defined as in Section 4.3.2. A change point is detected when $\Delta\text{BIC}(i) > 0$. If no change point is found in the current window, the window length is increased and the search is repeated. Once a maximum search window length is reached and no change is found, a change point is declared and the process is restarted. When a change point is found, a new search window is begun one vector after the detected change point.

To help minimise the cost of computing the BIC statistics at every point, a faster Hotelling's T^2 test is first used to identify the potential change point in a search window[14]. The full BIC statistic is then computed for the point with the maximum Hotelling's T^2 value in the window.

After the above process is run on the entire audio sequence, a second-pass BIC test is run on each detected change point to determine if adjacent segments should be merged. This second-pass mainly helps in eliminating very short segments and artificial change points due to reaching the maximum search window length.

When advert detection is used (as discussed in Section 4.1), detected advert regions are skipped during the change point detection.

Based on experimentation, the following settings are used for the change point detection algorithm: An initial search window size of 100 frames, a search window increment of 50 frames, a maximum search window size of 1500 frames, and α=1.0.

The segments are then classified as speech or non-speech using a GMM based maximum likelihood classifier. Five 128 mixture diagonal covariance GMMs are built for `Speech`, `Speech+Music`, `Speech+Other`, `Music` and `Other`. Any segments labelled as `Music` or `Other` are discarded before clustering. Further details can be found in [7].

5.2. Clustering

The speech segments are next clustered into speaker-homogeneous groups using a hierarchical agglomerative clustering approach[15] with the following steps:

0. Initialise leaf clusters of tree with speech segments.
1. Compute pair-wise distances between each cluster using a tied-mixture based generalised likelihood ratio distance.
2. Merge closest clusters.
3. Update distances of remaining clusters to new cluster.
4. Iterate steps 1-3 until stopping criteria is met.

The distance between clusters is :

$$d_{TGMM}(x,y) = -\log \frac{p(z|\lambda_z)}{p(x|\lambda_x)p(y|\lambda_y)} \qquad (5)$$

where x and y are the data from two different clusters, z is the union of x and y, and $p(x|\lambda_x)$ is the likelihood of data x given the pdf model λ_x for data x. The pdf model used is a tied-mixture model where the basis densities are estimated from the entire set of speech segments and the weights are estimated for each segment. Advantages of this model are the per-frame likelihoods to the basis densities need only be computed once and the weights for merged clusters are computed as a simple averaging of counts. The BIC criterion for this case is:

$$\Delta\text{BIC}_{TGMM} = d_{TGMM}(x,y) - \alpha\left(\frac{1}{2}M\log N\right) \qquad (6)$$

where M is the number of basis densities (and hence the number of free parameters) and N the total number of feature vectors. The clustering is stopped when $\Delta\text{BIC}_{TGMM} > 0$. Again, the penalty weight α, was set to 1.0, whilst M was 128.

5.2.1. Speech activity detection (SAD) gating

The purpose of this step is to detect and remove short bits of silence from the segments which can give rise to false-alarm errors in the scoring. A simple energy-based speech activity detector is run on the entire audio file to produce time marks of silence regions. Strictly speaking this is just an activity detector, since only the energy of the signal is used. The detected silence regions are gated out of the final segments prior to gender classification.

5.2.2. Gender classification

Gender classification is applied to the final speaker clusters. A GMM-based maximum likelihood classifier is applied to the aggregation of all data from a cluster. Using this approach, rather than classifying each segment independently, ensures a single gender label for all segments from a single speaker label. The gender classifier uses adapted GMM models[16] trained using data from the 1996 Hub4 training data set. A maximum of 2 hours of speech with High, Medium and Low quality labels for both male and female speakers (up to 6 hours of speech per gender) is used to train a 1024 mixture base GMM. The male and female speech is then used to adapt male and female models, respectively, from the base model. Using adapted models allows for a fast scoring technique[16] that significantly reduces the required computation. The gender classification error rate is 2.2% on the bndidev03 diarisation development data and 1.2% on the bneval03 data.

6. Building hybrid systems

A three-stage diarisation architecture was defined where each stage could be one of several options including the 'PERFECT' case, derived from the manually generated reference file. The stages are broken down as follows:

1. Advert Removal

NONE The advert removal stage was bypassed and the whole shows were passed on untouched.

CU_EVAL, CU_TDT4 Automatic advert detection as described in Section 4.1.

PERFECT All regions marked as adverts in the reference UTF file were removed.

The output from this stage consisted of a list of portions of audio for each show which were left after the advert removal stage.

2. Segmentation

CUED The CUED segmentation system described in Section 4.2. This included the music-removal and gender-relabelling stages.

MIT The MIT-LL segmentation system described in Section 5.1. Segment-level gender labels were additionally provided by MIT-LL, whilst bandwidth labels were automatically generated by CUED using the wide and narrow band models from the CUED segmenter in a GMM.

PERFECT Manual segmentation derived from the diarisation reference file. The times and gender of each segment were taken, but the speaker-id was ignored. Bandwidth labels were added automatically by CUED.

The output from this stage was a list of segments with bandwidth and gender labels.

3. Clustering

CUED The CUED clustering with BIC-local stopping criterion as described in Section 4.3.

MIT The MIT-LL clustering as described in Section 5.2. This included the final speech-activity-detection gating and the cluster-based gender-labelling stages.

PERFECT Cluster labels are assigned so as to maximise the overlap with the reference speakers in the diarisation reference file. The success of this obviously depends on the segment-purity of the preceding segmentation stage.

The results from running all combinations of this hybrid 'Plug-and-Play' system on the diarisation development (bndidev03) data are given in Table 2 and illustrated in Figure 5.

				Adverts excluded			Adverts as FA	
ADV	SEG	CLU	GE	MS	FA	DIA	FA	DIA
NO NE	CU	CU	1.9	0.2	9.1	25.54	29.7	46.14
	CU	MIT	2.1	2.5	5.3	24.23	24.7	43.60
	CU	PER	0.4	0.2	9.1	11.60	29.7	32.20
	MIT	CU	2.5	0.4	9.3	27.67	31.5	49.91
	MIT	MIT	2.2	2.7	5.6	24.46	26.8	45.68
	MIT	PER	0.6	0.4	9.3	11.67	31.5	33.91
CU - EV AL	CU	CU	2.0	0.6	9.1	25.89	23.5	40.34
	CU	MIT	1.7	2.9	5.3	24.92	18.7	38.38
	CU	PER	0.5	0.6	9.1	11.65	23.5	26.11
	MIT	CU	2.5	1.4	9.2	26.87	25.1	42.81
	MIT	MIT	2.3	3.7	5.6	25.93	20.6	40.96
	MIT	PER	0.6	1.4	9.2	12.54	25.1	28.49
CU - TD T4	CU	CU	2.3	1.0	8.9	27.03	12.6	30.80
	CU	MIT	1.8	3.4	5.1	26.67	8.2	29.80
	CU	PER	0.8	1.0	8.9	12.69	12.6	16.46
	MIT	CU	2.4	1.8	8.9	28.37	12.7	32.18
	MIT	MIT	1.7	4.1	5.3	25.02	8.5	28.26
	MIT	PER	0.6	1.8	8.9	12.67	12.7	16.48
PE RF	CU	CU	2.0	0.3	9.0	25.03	10.0	26.06
	CU	MIT	2.4	2.7	5.2	27.18	5.8	27.73
	CU	PER	0.6	0.3	9.0	11.93	10.0	12.96
	MIT	CU	2.5	0.9	9.3	26.12	10.3	27.12
	MIT	MIT	2.2	3.2	5.6	25.78	6.1	26.30
	MIT	PER	0.6	0.9	9.3	12.12	10.3	13.12
PE RF	PER	CU	0.0	0.0	0.0	18.71	0.0	18.73
	PER	MIT	2.3	2.5	0.0	17.55	0.0	17.57
	PER	PER	0.0	0.0	0.0	0.00	0.0	0.00

Table 2: Effect on diarisation score of using CUED, MIT and PERFECT components for advert-removal, segmentation and clustering within the 'Plug-and-Play' hybrid system. Scores are given both for the primary metric (where adverts are excluded from scoring) and a secondary metric which effectively counts adverts as silence in scoring, giving rise to additional FA errors if they are in the hypothesis. Scores are given for gender confusion error (GE), miss (MS), false alarm (FA) and total diarisation score (DIA) and are on the bndidev03 data.

6.1. Analysis of components

Advert Removal The average diarisation score for the different levels of advert removal are 20.9%, 21.3%, 22.1% and 21.4% for no-removal, CUED_EVAL, CUED_TDT4 and perfect removal, when excluding adverts from scoring. As the amount of automatic advert removal increases the diarisation score increases due mainly to the increase in false alarm rate due to incorrectly removed news. However, when adverts are treated as silence in scoring, the scores become 41.9%, 36.2%, 25.7% and 22.2% respectively, showing that the automatic advert removal helps, cutting the error rate by almost 40% relative.

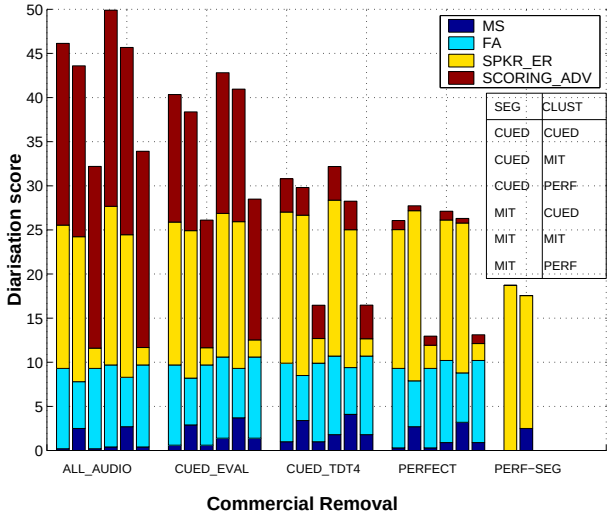

Figure 5: Results from the 'Plug-and-Play' hybrid diarisation system on the development (bndidev03) data. The inset table shows the ordering of the bars in each group of six.

Segmentation The segmentations are of similar quality, the average perfect clustering score being 12.0% on the CUED segmentation and 12.3% on the MIT segmentation, whilst the averages when using automatic clustering are 25.8% and 26.3% respectively.

Clustering The CUED and MIT clustering components are both quite robust to changes in input segmentation, with a variance of 1.1% and 1.0% respectively over all the segmentations. The MIT clustering component slightly outperforms the CUED system, the average score across the different segmentations being 25.5% and 26.6% respectively. It is interesting to note that the CUED clusterer always does better on the CUED segmentations, whereas the MIT clusterer sometimes does better on CUED segmentations and sometimes on MIT segmentations.

6.2. The best system

The best diarisation results from fully-automated runs are given in Table 3. For both diarisation scores the best system is a hybrid, using some components from CUED and some from MIT.

Adv	Seg	Clust	Score	Metric
NONE	CUED	MIT	24.23	Primary
CUED_TDT4	MIT	MIT	28.26	Secondary

Table 3: Best diarisation scores from a fully-automated hybrid system. The primary score excludes adverts in scoring, whilst the secondary score treats them as silence regions.

7. Conclusions

This paper has described the CUED December 2003 diarisation system, introducing a new 2-way splitting process with two new possible stopping criteria within the clustering component. This system gives a diarisation score 7 → 8% absolute better

than the CUED RT-03s evaluation system on both the development and evaluation data, and is considerably more robust to changes in segmentation.

The MIT-LL RT-03s diarisation system was also described, and a new hybrid 'Plug and Play' system was developed to allow the benefits of both the CUED and MIT-LL systems to be exploited in a single system. Analysis showed that on average the best performance came from using the CUED advert detection (when adverts were not excluded from scoring) and segmentation stages, whereas the MIT-LL clustering generally performed best. The lowest diarisation error rate whether adverts were excluded from scoring or not, came from a hybrid system, outperforming the individual systems from either site.

Future work will look at removing the 'Plug-and-Play' method's restriction on the diarisation systems having a common architecture, by combining the outputs from different diarisation systems directly using a cluster-voting scheme.[17] This could potentially allow information from many different systems (including those that do segmentation and clustering in a single stage) to be integrated to try to improve diarisation performance further.

8. References

[1] 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.

[2] A. Martin, "RT-02 Metadata Scoring," in *Proc. 2002 Rich Transcription Workshop (RT-02)*, Vienna, VA, May 2002.

[3] A. Martin and M. Przybocki, "Speaker Recognition in a Multi-Speaker Environment," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, September 2001, vol. 2, pp. 787–790.

[4] NIST, "Benchmark Tests : Speaker Recognition Evaluations," http://www.nist.gov/speech/tests/spk/.

[5] 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.

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

[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 Department, October 2003.

[8] S.E. Johnson, P. Jourlin, K. Spärck Jones, and P.C. Woodland, "Spoken Document Retrieval for TREC-8 at Cambridge University," in *The Eighth Text REtrieval Conference (TREC-8)*, E.M. Voorhees and D.K. Harman, Eds., Department of Commerce, NIST, Gaithersburg, MD, 2000, number SP 500-246, pp. 197–206.

[9] S.E. Johnson and P.C. Woodland, "A Method for Direct Audio Search with Applications to Indexing and Retrieval," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000, vol. 3, pp. 1427–1430.

[10] F. Bimbot and L. Mathan, "Text-Free Speaker Recognition using an Arithmetic Harmonic Sphericity Measure," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, Berlin, Germany, September 1993, vol. 1, pp. 169–172.

[11] 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. Automatic Speech Recognition and Understanding Workshop (ASRU)*, St. Thomas, U.S. Virgin Islands, December 2003, pp. 105–110.

[12] S.S. Chen and P.S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, May 1998, vol. 2, pp. 645–648.

[13] S.S. Chen, E. Eide, M.J.F. Gales, R.A. Gopinath, D. Kanvesky, and P. Olsen, "Automatic Transcription of Broadcast News," *Speech Communication*, vol. 37, pp. 69–87, 2002.

[14] P. Zhan, S. Wegmann, and L. Gillick, "Dragon Systems' 1998 Broadcast News Transcription System For Mandarin," in *Proc. 1999 DARPA Broadcast News Workshop*, Herndon, VA, March 1999, pp. 183–186.

[15] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation of Speech Using Speaker Identification," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Adelaide, Australia, April 1994, vol. 1, pp. 161–164.

[16] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Mixture Models," *Digital Signal Processing*, vol. 10, pp. 181–202, 2000.

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

Many Cambridge University Engineering Department publications are available from
http://mi.eng.cam.ac.uk/reports
and those associated with the DARPA Effective, Affordable Reusable Speech-to-text (EARS) programme can be found via
http://mi.eng.cam.ac.uk/research/projects/EARS/references.html