Cluster Voting for Speaker Diarisation

S.E. Tranter

CUED/F-INFENG/TR-476

1st May 2004

Abstract:

It is often important to be able to automatically detect 'who spoke when' in audio data. The speaker diarisation task attempts to address this problem on Broadcast News data by defining an error rate which can be used to evaluate segmentations and their associated speaker labels. Many different methods exist to automatically generate such segmentations and it would be desirable if segmentations from different origins could be combined to produce a more accurate one. This paper introduces a *cluster voting* scheme which attempts to use information from more than one diarisation system to produce a new speaker segmentation with a lower diarisation error rate. The scheme first generates a set of possible segmentations which minimise a distance metric based on the diarisation error rate and then defines a method of picking the final output from this set. Experiments presented using two inputs confirm that the diarisation error rate can be reduced using this new method.

Cambridge University Engineering Department
Trumpington Street
Cambridge
CB2 1PZ
England
Email: `sej28@eng.cam.ac.uk`

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.

Contents

Abbreviations used

BIC	Bayesian Information Criterion
BN	Broadcast News
Clust	Clustering or Clusterer
CUED	Cambridge University Engineering Department
CVM	Cluster Voting Metric
CVOS	Cluster Voting Output Set
DER	Diarisation Error Rate = MS + FA + SPE (%)
FA	False Alarm component of diarisation score (%)
GMM	Gaussian Mixture Model
MFCC	Mel-Frequency Cepstral Coefficients
MIT-LL	MIT Lincoln Laboratory
MS	Missed Speech component of diarisation score (%)
PLP	Perceptual Linear Prediction
RT-03s	NIST Rich Transcription 2003 Spring Evaluation
RT-03f	NIST Rich Transcription 2003 Fall (Metadata) Evaluation
SPE	SPeaker Error component of diarisation score (%)
STT	Speech-To-Text transcription
Seg	Segmentation or Segmenter
TF	Narrowband (telephone bandwidth) Female
TM	Narrowband (telephone bandwidth) Male
WER	Word Error Rate
WF	Wideband Female
WM	Wideband Male

bndidev03	RT-03s English BN diarisation development data set (6 shows)
bneval03	RT-03s English BN STT evaluation data set (6 shows)

1 INTRODUCTION

It can be very useful to know what events are happening within an audio stream without necessarily needing to generate a full transcription of the audio. Labelling phenomenon such as the gender of the speaker or the bandwidth of the channel can of course be helpful to aid subsequent speech recognition, but providing other information, such as labelling which speaker is talking, what is going on in the background or whether music is being played can be beneficial in its own right.

Consider, for example, analysing a Broadcast News show. Knowing the speaker-ids could help understanding by allowing speakers to be tracked through debates; information retrieval or browsing by allowing the user to locate when a particular speaker was talking; or even summarisation by revealing when the main newsreader was talking. Knowing the location of other events such as station jingles, commercials or music can help reveal the broadcast structure and allow information-less portions to be discarded, hence saving processing time, storage space, browsing time and helping retrieval efficiency (Johnson, Jourlin, Spärck Jones and Woodland 2001).

The general process of marking what is happening in the audio data has been termed 'diarisation' or 'speaker diarisation' where the interest is confined to marking up 'who spoke when' in the audio. This encompasses both audio segmentation and (speaker) clustering in a single task, and was introduced in the 2003 Rich Transcription Spring (RT-03s) Evaluation (NIST 2003b) although is a natural extension of the RT-02 Broadcast News Metadata speaker segmentation task (Martin 2002) which in turn developed from the (multi-speaker) speaker segmentation task in the speaker recognition evaluations (NIST 2002, Martin and Przybocki 2001, NIST 2000+).

Many methods to perform diarisation automatically have been tried, for example (Ajmera and Wooters 2003, Gauvain and Barras 2003, Nguyen and Junqua 2003, Tranter and Reynolds 2004, Moraru, Meignier, Besacier, Bonastre and Magrin-Chagnolleau 2004). These often use complementary approaches and it would be interesting to see if improvements could be made by combining more than one of these systems together. The speech recognition community have found significant benefit from combining different recognition outputs using the ROVER (Fiscus 1997) voting scheme (see e.g. (Woodland, Hain, Johnson, Niesler, Tuerk, Whittaker and Young 1998, Evermann and Woodland 2000)) and similarly the speaker identification/verification community have found that combining information from different systems before generating the final acceptance/rejection decisions can improve performance (see e.g. (Reynolds, Andrews, Campbell, Navratil, Peskin, Adami, Jin, Klusacek, Abramson, Mihaescu, Godfrey, Jones and Xiang 2003, Campbell, Reynolds and Dunn 2003, Kinnunen, Hautamäki and Fränti 2003)).

Some attempts to integrate different speaker segmentation/clustering systems have been made for diarisation, for example the 'Plug and Play' method described in (Tranter and Reynolds 2004) which combines the discrete components of the CUED and MIT-LL diarisation systems, or the 'hybridization' or 'piped' CLIPS/LIA system of the ELISA consortium (Moraru, Meignier, Besacier, Bonastre and Magrin-Chagnolleau 2003, Moraru et al. 2004) which works in a similar way. Both these combined systems have been shown to produce a lower diarisation error rate on Broadcast News data than the separate constituent systems. However, these approaches tend to place some restrictions on the component diarisation systems, for example by requiring separate stages for segmentation and clustering.

The ELISA consortium have also tried two different 'fusion' systems which 'merge' different segmentations before a final re-segmentation stage. The merging process described in (Moraru et al. 2004) effectively makes a new set of speaker labels based on the concatenation of all the input labels for any given frame and thus forms a new speaker segment every time the speaker of one of the inputs changes. This can help pick up speaker boundaries using complementary methods which may have been impossible to detect using a single segmentation system, but will tend to overgenerate said boundaries as the number of input segmentations is increased. However this effect is limited because short segments are generally removed during the final re-segmentation process.

The merging process described in (Moraru et al. 2003) first matches the two input segmentations using the metric used in the 2002 NIST speaker segmentation task (NIST 2002). Sections which agree are kept and can be used to train potentially more accurate speaker models, whilst the rest of the data is re-segmented. This technique showed improvements in performance over the component systems for the switchboard cellular and meetings data parts of the NIST 2002 speaker segmentation evaluation, although the results were not quite as convincing on the Broadcast News data, and the system restricted the two input segmentations to have the same number of speakers.

It would be interesting to see if trying to combine different system outputs directly just before scoring, such as is successfully used in both speech recognition and speaker verification could also be beneficial in the speaker diarisation task. This not only places no restriction on the diarisation systems used to generate the outputs but also allows the information to be combined *after* the stopping criteria (which can often be the most influential factor in speaker clustering systems) have been applied. This task poses different problems to its recognition equivalent, since the speaker labels, unlike transcribed words, are relative to each other and have no absolute meaning, so a method of associating the corresponding labels of the two diarisation outputs must first be made before any voting can be performed.

This paper describes a *cluster voting* scheme for combining speaker diarisation systems. It uses the diarisation speaker mapping and error rate scoring procedures to produce a new set of possible diarisation outputs from which the final outcome is chosen using a model-selection technique. The paper is arranged as follows. Section 2 defines the diarisation task, data and scoring in more detail, section 3 describes the theoretical framework of the cluster voting scheme and section 4 describes the methods of selecting the final output. Two experiments are then described, the first in section 5 uses two similar diarisation systems for inputs to show how the voting works in detail, whereas the second in section 6 combines the best two diarisation systems from CUED to try to reduce the diarisation error rate further. Finally conclusions are given in section 7.

2 The Diarisation Task

2.1 What is Diarisation?

Diarisation involves splitting up an audio stream into its main sources. In its general form this can include labelling events in the audio, such as music, speech, noise, laughter, background events, location of adverts etc. and associated properties or attributes (type of music, speaker-id, gender of speaker, source of noise etc.). For the purposes of the NIST RT-03s diarisation evaluation(NIST 2003b), the task was constrained to consider only the identification of speakers and (optionally) their gender.

The diarisation task considered here can thus be thought of as labelling ”who spoke when” in some audio, and consists of automatically producing a series of start/end time marks with associated speaker (and optionally gender) labels. Since the scoring is purely time-based and makes no use of the words spoken, most diarisation systems are built solely based on audio characteristics (see e.g. (Ajmera and Wooters 2003, Gauvain and Barras 2003, Nguyen and Junqua 2003, Tranter and Reynolds 2004)), although there is no reason why any other automatically derived information, such as from Speech-To-Text (STT) systems could not be used.

2.2 The Broadcast News Data for Diarisation

The experiments reported in this paper use the RT-03s English Broadcast News (BN) data. There are 2 main data sets marked up for diarisation,[1]

bndidev03 - the 6 development shows for the RT-03s diarisation evaluation, which were broadcast between 6th October and 19th December 2000.

bneval03 - the 6 shows used in the RT-03s STT evaluation. This is also the 3 shows used in the RT-03s diarisation evaluation and the 3 shows used for development for the RT-03f metadata evaluation. These shows were broadcast between 6th and 28th February 2001.

The data originates from US television and radio shows and each set consists of one 30 minute long episode from 6 different broadcasters, 2 radio, namely Voice of America English News (VOA_ENG) and PRI The World (PRI_TWD); and four TV 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).

The data is very challenging for diarisation since it contains many different types of genre, for example adverts, station jingles, announcements from within a studio, reports from in the field, discussions of important issues between several people; and many different types of acoustic condition, for example the existence of background noise or music, or different bandwidths from different locations

The speaker diarisation task is made considerably harder by not knowing the number of speakers in advance, and the wide variety in the amount of time the speakers spoke for, ranging from a single word for some interviewees to over a thousand for some anchor presenters. A distribution of the loquacity of the speakers in the bndidev03 and bneval03 data sets is given in Figure 1.

An additional complication is the fact that commercial breaks are included within the broadcast shows. Although these regions are currently excluded from scoring for diarisation they can still detrimentally affect performance by for example interacting with the target data in clustering. For this reason, it may be desirable to attempt to remove the adverts automatically before recognition. This can be done with a moderate degree of success if contemporaneous (unannotated) data is available (see e.g. (Johnson and Woodland 2000)), but this paper does not consider using such a system so as to focus on the main speaker diarisation task. Finally, this data also contains some portions of overlapping speech, where more than one person is speaking simultaneously, although these regions were excluded in the primary evaluation metric.

[1] The exact composition of the datasets can be found in (Tranter, Yu, Reynolds, Evermann, Kim and Woodland 2003).

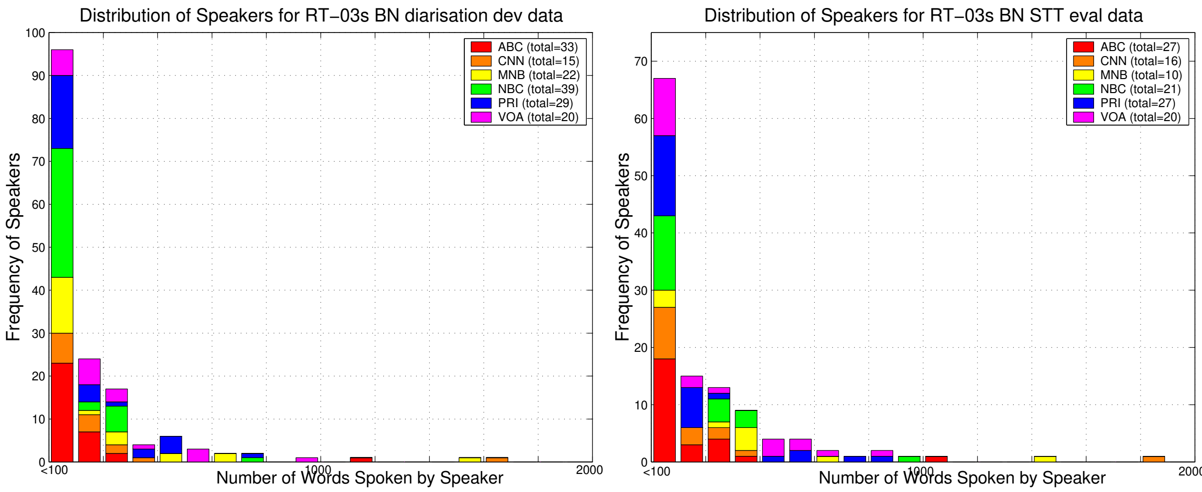

Figure 1: *Distribution of speakers for the bndidev03 and bneval03 data sets*

2.3 Diarisation Scoring

The rules for the diarisation component of the Rich Transcription Spring 2003 (RT-03s) evaluation are described in (NIST 2003b).

A reference file is generated from the word-level transcripts, giving 'ground-truth' time-marked speaker segments. A speaker turn is broken up into distinct speaker segments when either a new speaker starts talking, or the speaker pauses for more than a certain critical length of time, which was set at 0.3s. The word times which were used to derive these speaker segments were generated by the LDC by performing a forced alignment of the reference words in each given speaker turn.

In addition to the reference speaker segments, a list of regions to exclude from scoring is also provided. This corresponds to adverts in broadcast news shows, or speaker-attributable vocal noises such as cough, breath, lipsmack, sneeze and laughter. Further details of the reference generation process can be found in (NIST 2003a).

The performance of a system hypothesised speaker segment list is evaluated by first computing an optimal one-to-one mapping of reference speaker IDs to system output speaker IDs for each broadcast news show independently. This mapping is chosen so as to maximise the aggregation over all reference speakers of the time that is jointly attributed to both the reference and the (corresponding) mapped system output speaker.[2]

Speaker detection performance is 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:

$$\text{DER} = \frac{\sum_{allsegs} dur(seg) \cdot (\max(N_{Ref}(seg), N_{Sys}(seg)) - N_{Correct}(seg))}{\sum_{allsegs} dur(seg) \cdot N_{Ref}(seg)}$$

[2] This is computed over all regions of speech, including regions with overlapping speech.

where :

DER	is the total diarisation error rate
seg	is the longest continuous piece of audio for which the reference and hypothesised speakers do not change
$dur(seg)$	is the duration of the *seg*
$N_{Ref}(seg)$	is the number of reference speakers in the *seg*
$N_{Sys}(seg)$	is the number of hypothesised speakers in the *seg*
$N_{Correct}(seg)$	is the number of mapped reference speakers which match the hypothesised speakers

This formula allows the whole file to be evaluated, including regions of overlapping speech. For the primary evaluation score, where regions containing multiple simultaneous speakers are excluded, this formula reduces to[3]

$$\text{DER} = \frac{\sum_{allsegs} dur(seg) \cdot (H_{miss} + H_{fa} + H_{spe})}{\sum_{allsegs} dur(seg) \cdot H_{ref}}$$

where

$H_{miss} = 1$	iff speaker is in reference but not in hypothesis, else 0
$H_{fa} = 1$	iff speaker is in hypothesis but not in reference, else 0
$H_{spe} = 1$	iff mapped reference speaker does not equal hypothesis speaker, else 0
$H_{ref} = 1$	iff seg contains a reference speaker, else 0

Since the segments are time-weighted, this metric is biased towards getting the most prolific speakers correct. For example if the system incorrectly splits a 5-minute reference speaker into 2 equally-sized clusters, this gives a 50% higher error rate than missing 10 different speakers of 10s duration.

[3] Assuming systems do not output files containing overlapping speakers.

3 Cluster Voting

The aim of cluster voting is to take the output from different clustering schemes and try to form a new output which is better than any of the inputs. For some cases it may be that there is no difference between the strength of evidence supporting a number of alternatives for a given decision from the inputs to the cluster voting, and a decision must be taken some other way, for example, arbitrarily, based on prior knowledge or preferences, or using an external 'judge'. The latter option is preferred providing a suitable judge can be found, but there may be cases where one of the other strategies is sufficient. The cluster voting architecture is illustrated in Figure 2.

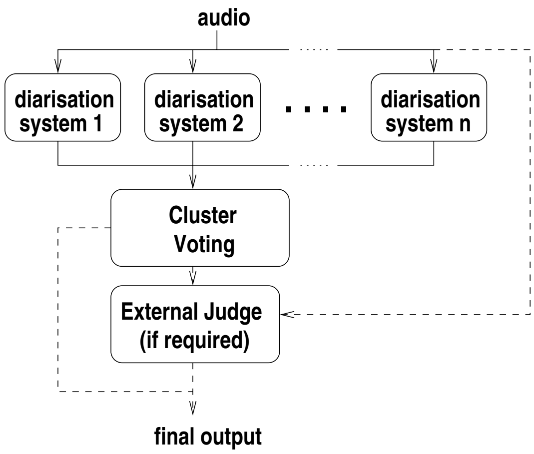

Figure 2: *Cluster Voting Architecture*

For this work only the two input case is considered for simplicity, but the same general principles can be extended to cover the case where more inputs are available to the cluster voting.

3.1 The Cluster Voting Metric (CVM)

In order to produce a new diarisation output from the diarisation inputs a metric must be defined which can be minimised. Since it is hoped that the output which minimises this metric will produce a lower overall diarisation error rate, it was felt it would be beneficial to use the actual metric used in the final diarisation scoring within the cluster voting minimisation.

Since we assume that no system output contains overlapping speakers, we can use the simplified version of the diarisation scoring metric as described in section 2.3 namely:

$$\text{DER} = \frac{\sum_{allsegs} dur(seg) \cdot (H_{miss} + H_{fa} + H_{spe})}{\sum_{allsegs} dur(seg) \cdot H_{ref}}$$

In this case we postulate a new clustering of the base segmentation and calculate the optimum speaker mapping from this to both inputs. Maximising the sum of the amount of overlap from the new clustering to both inputs gives the same result as minimising the sum of the DERs from the new clustering to both the inputs, thus the metric to *maximise* within the cluster voting stage is defined as the sum of the overlap in time between the postulated output clustering and the two inputs under the constraint of the optimum speaker mapping. We call this the *Cluster Voting Metric (CVM)*.

Different possible clusterings are then postulated and scored and the ones with the (joint) highest CVMs are listed as the possible outputs in the *Cluster Voting Output Set (CVOS)*.

For the two-input problem there will always be a set of possible clusterings, the Cluster Voting Output Set (CVOS),[4] that give the highest CVM, and this set will always include both of the inputs. One way to find all the members of this set is to generate all the possible output clusterings given the base segmentation and exhaustively search them to find the ones that give the maximum in the CVM. However, this is very computationally expensive so methods of reducing the search space must be sought.

Once the CVOS has been obtained, a method of choosing between the set members must be defined. This could be a simple strategy such as picking one at random, or labelling the conflicting segments with the same (or all different) speakers; or something more complicated, such as using information from a third input or an external judging scheme. The choice between CVOS alternatives only needs to be made for the subset of segments which have conflicting labels. Some segment labels will be fixed during the CVM maximisation so these can bypass the judging stage and be output directly from the cluster voting scheme.

3.2 Generating all Possible Output Clusterings

One way to ensure that all the members of the Cluster Voting Output Set (CVOS) are found during the cluster voting is to postulate all possible output clusterings for a given segmentation and then find those clusterings which maximise the Cluster Voting Metric (CVM). This is computationally expensive but steps can be taken to reduce the complexity to allow this to be done in some cases.

The first step is to form the base segmentation. If the inputs to the cluster voting come from different clusterings of a common segmentation then this step is unnecessary, but if the original segmentation is different for the different inputs a new base segmentation must be defined in which no speaker changes for either input. This is the same as is done in the diarisation scoring (see section 2.3). Once the base segmentation is known all the different possible output clusterings can be generated directly.

3.2.1 The Bell Combination Series

The series which represents all possible distinct output clusterings of N separate segments is a very interesting one, called the Bell combination series.[5]

N	Possible Output Clusterings	
	Number	Sequences
1	1	(1)
2	2	(1 1); (1 2)
3	5	(1 1 1); (1 1 2); (1 2 1); (1 2 2); (1 2 3)
4	15	(1 1 1 1); (1 1 1 2); (1 1 2 1); (1 1 2 2); (1 1 2 3)
		(1 2 1 1); (1 2 1 2); (1 2 1 3); (1 2 2 1); (1 2 2 2)
		(1 2 2 3); (1 2 3 1); (1 2 3 2); (1 2 3 3); (1 2 3 4)

Table 1: *The Bell series for generating all possible output clusterings for N distinct segments.*

At first glance it appears to be an $N!$ choice, since the first cluster is given an id of 1, the second can be 1 or 2, the third 1,2 or 3 etc. but in fact several of these groupings are equivalent and the number of *different* clusterings is less. Table 1 gives the possible sequences for the cases of N=1 to N=4. The first case is trivial. Two segments can be grouped as either both the same or both different, leading to two possibilities. Three segments generates all the sequences expected from the factorial sequence *except* for `(1 1 3)` which is equivalent to `(1 1 2)` in that in both cases the first two segments are given the

[4]Since the optimal mapping in the diarisation metric can include 'NULL's it is possible that certain segments can be assigned an arbitrary speaker-id without detrimentally affecting the CVM. This is why for some cases there can be many different possible clusterings in the CVOS.

[5]Further information about the Bell series is given in Appendix A.

same cluster-id but the third is given a different one. This leads to 5 possibilities instead of the 6 that would be given from the factorial when N=3. Similarly, when N=4, any sequences which start `(1 1 3)` can be ruled out for the same reason, as can {`(1 1 1 3), (1 1 1 4), (1 1 2 4), (1 2 1 4), (1 2 2 4)`} leaving 15 in total.

Table 2 gives the number of possible outputs for N distinct input segments, B(N), as N increases, and compares this with N!. Although the sequence does grow in an factorial-like fashion, its actual value becomes a smaller proportion of the factorial case as N increases, leading to approximately 4 million, instead of 479 million when N is 12, making the generation of every possible combination feasible for values of N up to around 12.

N	N!	B(N)
0	1	1
1	1	1
2	2	2
3	6	5
4	24	15
5	120	52
6	720	203
7	5,040	877
8	40,320	4,140
9	362,880	21,147
10	3,628,800	115,975
11	39,916,800	678,570
12	479,001,600	4,213,597
13	6,227,020,800	27,644,437

Table 2: *The number of possible partitions, B(N), when generating all output clusterings for N distinct segments.*

3.3 Increasing Computational Efficiency

Since the generation of the possible clustering outputs is very computationally expensive, ways of reducing the complexity need to be sought.

3.3.1 Base Segmentation

By defining a common base segmentation as discussed in section 3.2 and using the knowledge that none of the inputs to the cluster voting have overlapping speakers, the calculation of the amount of overlapping time between the different clusterings can be simplified. Each segment can be represented by a unique id and duration. When comparing two clusterings only the segments with matching ids need to be compared and they contribute the duration to the score if the speaker ids match and zero otherwise.

3.3.2 Resegmentation

The data can be resegmented by grouping together segments which contain a common speaker id for each of the inputs in turn. There is no need for these segments to be adjacent in time, so this is more powerful than a traditional 'merging' of segments with the same speaker. For example suppose the base segments number 1 and 3 are both labelled as SPKR1 in the first input, and are both labelled as SPKR2 in the second input then a new *'resegment'* can be formed which combines base segments 1 and 3 into the same segment. No decision is made about the speaker id of these segments at this time other than to force them to have the same speaker id in the final output. An example of this process is given in Table 3.

Base-Seg-id	Input-1	Input-2	Resegmentation
base1	A1	B2	R1
base2	A1	B1	R2
base3	A2	B1	R3
base4	A1	B2	R1
base5	A3	B3	R4
base6	A2	B1	R3
base7	A3	B3	R4

Table 3: *Example of the resegmentation process. Base segments base1 and base4 are clustered together in both inputs and so can be joined to form a single resegment. As can base3 and base6; and base5 and base7. This means the number of 'segments' to be dealt with has been reduced from 7 (base-segments) to 4 (resegments)*

3.3.3 Removing Non-Conflicting Segments

Any resegments which are unique, in that there is no segment in any input which has the same speaker id as the segment in the resegment but which itself not in the resegment, can be immediately given a unique output speaker id and output directly. This is because the CVM maximisation will always give a separate (non-overlapping) speaker id for such a resegment. In the example in Table 3 this applies to R4, since its constituent base segments (base5 and base7) have input speaker ids A3 and B3, neither of which is found in any other segment in the respective inputs. This reduces the number of 'segments' to be dealt with after resegmentation from 4 to 3 in this case.

3.3.4 Generating Independent Supergroups of Resegments

The speaker ids of some groups of segments can be independent of other groups of segments. For example, suppose the clustering has been done gender and bandwidth dependently. For this case there will be no possibility of overlap between the narrowband speakers and wideband speakers, or male and female speakers, so there will be a minimum of four independent *'supergroups'* of resegments. Treating these groups independently in the subsequent calculations does not reduce the number of resegments, but does dramatically reduce the number of combinations which must be searched, since it is known that resegments in one supergroup will never cluster with those in a different one.

These independent supergroups can be found automatically simply by tracing through the speakers of the resegments in turn using the following algorithm.

```
foreach resegment
   next if already assigned resegment to a supergroup
   push resegment to the todo list
   while the todo list is not empty
       pop the todo list to give the next resegment
       push the resegment onto the current-group list
       foreach input
          find speaker-id of the resegment in the input
          add resegments which have the same input speaker-id to the todo list
       end
   end
   start a new current-group list
end
```

An example of this process taken from the ABC show for example 2 described in section 6 is given in Table 4.

Resegment-id	input1	input2	group
R0	A-CWF4	B-CWF8	G0
R1	A-CWF4	B-CWF9	G0
R2	A-CWM2	B-CWM4	G1
R3	A-CWM2	B-CWM7	G1
R4	A-CWM2	B-CWM8	G1
R5	A-CWM6	B-CWM2	G2
R6	A-CWM7	B-CWM2	G2
R7	A-CWM8	B-CWM2	G2
R8	A-CWM9	B-CWM2	G2
R9	A-CWF4	B-CWF10	G0

Table 4: *Example of forming independent resegment super-groups. Here three groups are formed from the 10 input resegments. G0 corresponds to the inputs {A-CWF4,B-CWF8,B-CWF9,B-CWF10}, G1 corresponds to the inputs {A-CWM2,B-CWM4,B-CWM7,B-CWM8} and G2 corresponds to the inputs {A-CWM6,A-CWM7,A-CWM8,A-CWM9,B-CWM2}. For this case the complexity has been reduced from B(10)=115,975 to B(3)+B(3)+B(4)=25*

3.3.5 *Example of Complexity Reduction*

Taking the two inputs for example 2 as described in section 6 for the ABC show, the reduction in complexity by implementing each of the above strategies is shown in Table 5.

Number of segs in each input	127	B(127)=too many!
Number of resegs after resegmentation	22	B(22)=4.5067×10^{15}
Number of resegs after outputing unique cases	10	B(10)=115975
Definition of supergroups	{0,1,9}, {2,3,4}, {5,6,7,8}	25

Table 5: *Reduction in number of possible output clusterings for a 2-input example using the techniques discussed in section 3.3*

3.4 Using the Speaker Mapping between the Inputs

The optimal speaker mapping between the inputs can be used to restrict the number of possible output clusterings which need to be generated. This is because any output clustering which directly contradicts this mapping (such as assigns the same speaker-id to segments which are mapped to speakers in one input which are in turn themselves mapped to *competing* speakers in the other input), will never give the optimal CVM score and thus will never appear in the CVOS. This means it is not necessary to generate all possible output clusterings for the base segmentation.

4 Judging Possible Alternatives

Given that the Cluster Voting Output Set (CVOS) contains more than one possible output clustering which maximises the CVM, a method of choosing one possibility from this set must be defined. This could be done using a very simple strategy, such as always giving the resegments different speaker-ids; using confidence scores to weight alternatives; or exploiting additional information, for example from another input clustering, or from an external judge.

For the two-input case, since both inputs will always be in the CVOS, using confidence scores to weight one input over the other will simply mean the final output is identical to the input with the highest weight. This is not inevitable when there are more than two inputs, but since this work only considers the two-input case, this option is not discussed further.

Section 4.1 gives some examples of simple strategies to pick between alternatives, whilst section 4.2 discusses using an external judge based on the Bayes Information Criterion (BIC). Experiments which investigate the success of the strategies in question are reported in sections 5 and 6.

4.1 Simple Strategies

The following simple strategies can easily be used to pick between alternatives in the CVOS.

- Omit all conflicts from the final output (miss)
- Assign a single speaker-id for all the segments in a supergroup (same)
- Assign each resegment a different speaker id (diff)
- Pick one alternative from the CVOS at random (random)
- Use a confidence score to weight the inputs (N/A for 2-input case)

Results using these strategies are given in sections 5 and 6.

4.2 Using BIC to Decide Between Alternatives

4.2.1 Standard BIC

The standard Bayes Information Criterion (BIC) represents a way of selecting between a set of models by calculating a log likelihood of the data which is then penalised in proportion to the number of parameters in the model.(Chen and Gopalakrishnan 1998, Chen, Eide, Gales, Gopinath, Kanvesky and Olsen 2002)

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

where $\#M$ is the number of free parameters, N the number of data points and α the tuning parameter, usually set to 1.

If K clusters of the data are each modelled using a full Gaussian of dimension d which has N_i frames and a covariance S_i then the BIC formula becomes:[6]

$$\text{BIC} = -\frac{1}{2}\left(\left[\sum_{i=1}^{K} N_i \log(|S_i|)\right] + Nd(1+\log(2\pi)) + \alpha K\left(d + \frac{d(d+1)}{2}\right)\log N\right)$$

The best model-set can thus be chosen from a list of alternatives by finding the model-set which maximises the BIC term. This is the same as minimising

[6] See (Tranter and Reynolds 2004) for a more detailed derivation.

$$\text{BMIN} = \left[\sum_{i=1}^{K} N_i \log(|S_i|)\right] + \alpha K \left(d + \frac{d(d+1)}{2}\right) \log N \tag{2}$$

Thus if a single Gaussian is built for each speaker-id in a given output clustering, the BMIN value can be calculated for this output clustering, and the final output will be given by the clustering in the CVOS which gives the lowest BMIN value. Results from using this scheme are given in section 6.

4.2.2 Equal-Parameter BIC

An alternative implementation of a BIC-based strategy, introduced in (Ajmera, Bourlard and Lapidot 2002, Ajmera, McCowan and Bourlard 2004) and discussed within the diarisation framework in (Ajmera and Wooters 2003), removes the need for the tunable parameter, α, by making sure each model-set being compared always contains the same number of parameters, thus the '$\alpha \# M \log N$' term in equation 1 is independent of the model-set and can be removed from the minimisation. The choice of model-set thus reduces to that which gives the highest likelihood of the data.

To visualise this how this works for a simple case, consider whether to model a set of segments as a single cluster or as two clusters. The single-cluster case can be modelled by a 2-mixture GMM, whereas the two-cluster case can be modelled by a single Gaussian (or 1-mixture GMM) for each of the two clusters. In this way the total number of parameters used to model the data is the same in both cases.[7]

Suppose the data really comes from two distinct speakers, then the 2-mixture GMM will form two Gaussians which will almost certainly correspond to the two distinct speakers. However, supposing the mixture weights in the GMM are both 0.5, then the likelihood of all the data over the 2-mixture GMM will be roughly half that of the likelihood of running each speaker over its own Gaussian. Thus for the case of two distinct speakers, the two-cluster case is preferred.

Now suppose the data really all comes from the same speaker, and thus the distributions of the two-cluster case are almost identical to the one-cluster case. This means that both Gaussians in the 1-mixture GMMs will effectively model the data as a single Gaussian, but the 2-mixture GMM will be able to model the data with two Gaussians and thus will produce a higher likelihood. Therefore the one-cluster case will be preferred when the data really does come from the same speaker. Results from using this scheme are given in section 6.

[7] The 1 extra parameter introduced by the free mixture weight for the two-mixture GMM is ignored here.

5 Example 1 : Cluster Voting on Two Very Similar Systems

The first experiment with the cluster voting scheme took the output from two diarisation systems that were very similar for the two inputs. The CUED RT-03s base segmentation was used for each case and the clustering used a top-down two-way split with the arithmetic harmonic sphericity distance measure(Bimbot and Mathan 1993). Both clusterings used a BIC-based stopping criterion but the exact formulation and value of the α parameter was slightly different for each case, leading to small differences in the output.

The performance of the systems was similar, giving 25.12% and 25.21% diarisation error rate (DER) on the bneval03 data. More details about the segmentation are given in section 5.1, and the clusterings in section 5.2 and the results of the cluster voting scheme are given in section 5.3.

5.1 The Base Segmentation

The CUED RT-03s segmentation on the bneval03 data was taken as the base segmentation. This gave automatically produced segment labels which contained a start and end time, a gender (male or female) and a bandwidth (wideband or narrowband). The breakdown of the segment distribution by show is given in Table 6. Further details about the segmenter can be found in (Tranter et al. 2003).

Show	Number of Segments				
	WM	WF	TM	TF	TOTAL
20010206+1830+1900+ABC+WNT	70	56	0	1	127
20010217+1000+1030+VOA+ENG	83	31	3	12	129
20010220+2000+2100+PRI+TWD	68	44	30	2	144
20010221+1830+1900+NBC+NNW	127	37	14	2	180
20010225+0900+0930+CNN+HDL	59	97	4	1	161
20010228+2100+2200+MNB+NBW	98	20	10	0	128
TOTAL	505	285	61	18	869

Table 6: *CUED RT-03s segmentation: Segment distribution over gender and bandwidth for each show in bneval03*

5.2 The Input Diarisation Systems

The CUED December 2003 clustering system(Tranter and Reynolds 2004) was run over the base segmentation. The clustering is done bandwidth and gender dependently. The clustering algorithm uses a binary top-down splitting procedure. For each parent node two children nodes are formed and the segments are moved between the children nodes under a certain distance metric until either every segment prefers to be in its current child node, or the maximum number of iterations is reached. Stopping criteria are then applied to the possible split to decide whether the split should go ahead or not. This process is repeated until no further nodes can be split.

Two versions of the clusterer were run which differed only slightly in the definition of stopping criteria. The data was represented by the correlation matrix of 13-dimensional static only PLP coefficients, and the distance metric used was the arithmetic harmonic sphericity(Bimbot and Mathan 1993). Both schemes used the BIC-based stopping criterion defined in (Tranter and Reynolds 2004), but the first used the 'local' BIC scheme, where the number of samples in the penalty term is the number of frames in the *parent cluster*, whereas the second used the 'global' BIC scheme which used the total number of frames in *all the data* instead. The values of the α tuning parameter were chosen from optimal performance on the bndidev03 data, and were 7.25 and 6.25 respectively. More details can be found in (Tranter and Reynolds 2004).

5.3 Cluster Voting Experimental Results

The two inputs to the cluster voting scheme were very similar since they had been generated using the same method with only a slight variation in stopping criterion. The cluster voting resegmented the data using the method described in section 3.3.2 and non-conflicting segments were output directly with a new speaker-id as described in section 3.3.3. The remaining resegments were divided into supergroups as described in section 3.3.4. The number of re-segments, supergroups and the breakdown of the supergroups is given in Table 7.

Show	# Resegments		SuperGroups
	Total	Non-Conflicting	[ID/ Duration(s)/ # resegments]
ABC+WNT	18	18	-
VOA+ENG	18	12	[g0/30.23/2], [g1/170.86/2], [g2/43.46/2]
PRI+TWD	20	20	-
NBC+NNW	19	17	[g0/46.20/2]
CNN+HDL	25	21	[g0/239.75/2], [g1/88.70/2]
MNB+NBW	15	15	-

Table 7: *Example 1: Number of resegments and supergroups in the cluster voting*

Since there were so few differences between the two inputs, it was possible to exhaustively search all the combinations of the resegments and score the resulting outputs against the true reference to show the range of scores that could be achieved using the cluster voting scheme. Each supergroup was treated independently and since all the supergroups only contained two resegments, the possibilities were limited to a choice of three, namely:[8]

s Put both re-segments in the same cluster

d Put the 2 re-segments into 2 different clusters

m Do not output anything for the 2 re-segments (i.e. 'miss' them out)

The results for all possible combinations of output when scored against the true reference data are given in Table 8. For the cases of CNN and NBC, the DER can be reduced when compared to either input by allowing the segments which do not agree to be missed out of the output completely. However, for the VOA show, this is not the case and the option of missing all the disagreements out of the output gives a large increase in DER (from ˜20% to over 30%). Obviously missing all the conflicting segments out of the output becomes a very dangerous strategy when the difference between the inputs gets larger. Therefore this option will largely be ignored, although it has been noted that it could provide an improvement in performance for certain cases.

Table 9 shows a summary of the experimental results, giving the number of possible outputs, the best, worst, mean and median scores the scores from special cases such as picking one member of the CVOS at random, making the speaker ids all the same, making the speaker ids all different or missing all conflicting segments from the output.

Despite the inputs being so similar, it *is possible* to reduce the overall DER when compared to either input. The DERs for the inputs over the whole 6-show data set are 25.12% and 25.21%. The range of possible DERs from the postulated outputs are [24.95% - 25.38%] (= inputs $\pm$ 0.17%) when misses are not allowed and [24.68% - 27.44%] when misses are allowed. When misses are not allowed the mean score of the CVOS possibilities is the average of the two inputs, whilst the (lower) median and randomly-chosen output match the best of the two inputs.

[8]Note that for a supergroup with two resegments, the possible outputs in the CVOS consist of (a) give the two resegments the same spkr-id (b) give the two resegments a different spkr-id.

VOA								CNN			NBC	
g0	g1	g2	DER	g0	g1	g2	DER	g0	g1	DER	g0	DER
d	d	d	*19.94*	m	d	m	22.43	m	m	*36.46*	m	*31.74*
s	d	d	*20.34*	s	s	m	22.52	m	d	*36.86*	[s|d]	32.06 *i1,*i2
d	s	d	*20.76*	m	s	s	22.75	m	s	*36.90*		
d	d	s	20.78 *i1	m	s	m	23.26	[s|d]	m	*37.52*		
m	d	d	21.08	d	m	d	29.46	[s|d]	d	37.92 *i1		
s	s	d	21.17 *i2	s	m	d	29.86	[s|d]	s	37.96 *i2		
s	d	s	21.18	d	m	s	30.30					
d	d	m	21.29	m	m	d	30.60					
d	s	s	21.60	s	m	s	30.70					
s	d	m	21.69	d	m	m	30.81					
m	s	d	21.91	s	m	m	31.22					
m	d	s	21.92	m	m	s	31.44					
s	s	s	22.01	m	m	m	31.96					
d	s	m	22.12									

Table 8: *Example 1: Diarisation Error Rate (DER) for all possible outputs from the CVOS (including missing conflicts out). DERs are generated against the true reference data. (s) means the cluster group has the same cluster-id for the 2 resegments, (d) means they are different, and (m) means they are missed out from the output completely. *ix is the same as input x.*

Condition		Show			
		VOA	CNN	NBC	TOTAL (all 6)
Input 1		20.78	37.92	32.06	25.12
Input 2		21.17	37.96	32.06	25.21
No miss allowed	Num Poss	8	4	2	(14)
	Best	19.94	37.92	32.06	24.95
	Worst	22.01	37.96	32.06	25.38
	Mean	20.97	37.94	**32.06**	25.17
	Median*	20.78	**37.92**	**32.06**	25.12
	Random	20.78	**37.92**	**32.06**	25.12
	All the same	22.01	37.96	**32.06**	25.38
	All different	***19.94***	**37.92**	**32.06**	***24.95***
miss allowed	Num Poss	27	9	3	(39)
	Best	19.94	36.46	31.74	24.68
	Worst	31.96	37.96	32.06	27.44
	Mean	24.63	*37.45*	*31.95*	25.83
	Median	22.12	*37.52*	32.06	25.34
	Random	21.21	37.92	*31.74*	25.16
	All missed	31.96	**36.46**	**31.74**	27.17

Table 9: *Example 1: Summary of DERs for the possible cluster voting outputs. [* The lower median is taken here i.e. the n/2 th entry in the list of n possibilities]*

For this particular example, where there are few differences between the inputs, if we do not allow misses, the best results come when all the speaker ids are made different, and the worst results come when all the speaker ids are the made the same. However, this is a very simple example so no conclusions can be formed other than noting the possibility for improving (or degrading) performance using this method of cluster voting.

6 Example 2 : Cluster Voting on CUED's Two Best Systems

6.1 The Input Diarisation Systems

The two-way cluster voting experiment was repeated using the same first input (based on the local-BIC implementation) but a different second input. The latter was generated using the same base segmentation (as described in section 5.1) and the same clustering strategy but a different stopping criterion which was based on the relative decrease in node cost (as measured by the average distance from constituent segments to the node centre) for a given split. The parameter controlling this split was set at 0.825 which gave optimal performance on the bndidev03 data. Further details about this 'cost-based' system can be found in (Tranter and Reynolds 2004).

The standards of the two inputs are similar (namely 25.12% and 27.09%), the first giving 1.97% absolute lower DER across the 6-show bneval03 set, but the second input is in fact better for 5 out of the 6 shows but performs poorly on the remaining show due to a single bad decision to split a large cluster into two. It is unfortunate that this single decision produces such a large effect - which is partly down to the scoring procedure's bias towards prolific speakers, and partly due to the small number of shows in the test set.

6.2 The Cluster Voting Process

The number of re-segments and the breakdown of the supergroups is given in Table 10. This shows that the two inputs are considerably more different than in Example 1. The average number of of re-segments per show for which the inputs do not agree is 15, whilst the average number of supergroups formed from these re-segments is 5. However, it is still possible to score *all* the output combinations given by the cluster voting minimisation.

Show	# Resegments		SuperGroups
	Total	Unique	[ID / Duration(s) / # resegments]
ABC+WNT	22	12	[g0 / 127.3 / 3], [g1 / 90.56 / 3], [g2 / 232.87 / 4]
VOA+ENG	30	11	[g0 / 45.52 / 2], [g1 / 139.85 / 8], [g2 / 17.14 / 2] [g3 / 14.46 / 2], [g4 / 18.6 / 2], [g5 / 43.46 / 3]
PRI+TWD	26	13	[g0 / 43.19 / 3], [g1 / 265.98 / 3], [g2 / 82.51 / 3] [g3 / 34.66 / 2], [g4 / 1.28 / 2]
NBC+NNW	31	11	[g0 / 180.75 / 2], [g1 / 96.56 / 4], [g2 / 75.22 / 3], [g3 / 46.2 / 2] [g4 / 33.72 / 2], [g5 / 31.97 / 5], [g6 / 1.98 / 2]
CNN+HDL	33	13	[g0 / 239.75 / 2], [g1 / 49.92 / 2], [g2 / 34.62 / 6], [g3 / 139.37 / 2] [g4 / 70.69 / 3], [g5 / 46.9 / 3], [g6 / 3.89 / 2]
MNB+NBW	20	11	[g0 / 473.61 / 2], [g1 / 32.08 / 2], [g2 / 7.81 / 2] [g3 / 18.73 / 3]

Table 10: *Example 2: Number of resegments and supergroups when comparing the CUED's two best diarisation systems of December 2003. They were derived from using a BIC-based or a cost-based stopping criterion within the clusterer.*

6.3 Scoring All Possibilities from the Cluster Voting Output Set (CVOS)

The results from scoring all the members of the CVOS against the true reference data are summarised in Table 11. For this case all the data is given a speaker label (i.e. no 'misses' - where the data corresponding to an entire supergroup is omitted from the output - are allowed). Figure 3 gives a graphical representation of the key results.

Condition		ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
Input 1		32.03	20.78	21.40	32.06	37.92	10.74	25.12
Input 2		29.26	19.82	20.48	31.56	37.18	29.34	27.09
No misses allowed	Num Poss	128	8192	256	8192	8192	32	(24,992)
	Best	26.71	18.43	18.11	29.84	37.18	10.74	22.79
	Worst	34.58	22.48	23.56	33.78	37.92	29.34	29.44
	Mean	30.56	20.45	20.89	31.67	37.55	20.04	(26.09)
	Median*	30.30	20.45	20.74	31.66	**37.18**	**10.74**	*24.51*
	Random	*28.30*	*19.76*	23.18	***29.84***	**37.18**	**10.74**	*24.30*
	All the same	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
	All different	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02

Table 11: *Example 2: Summary of DERs for the possible cluster voting outputs when 'misses' are not allowed. [* The lower median is taken here i.e. the n/2 th entry in the list of n possibilities.]*

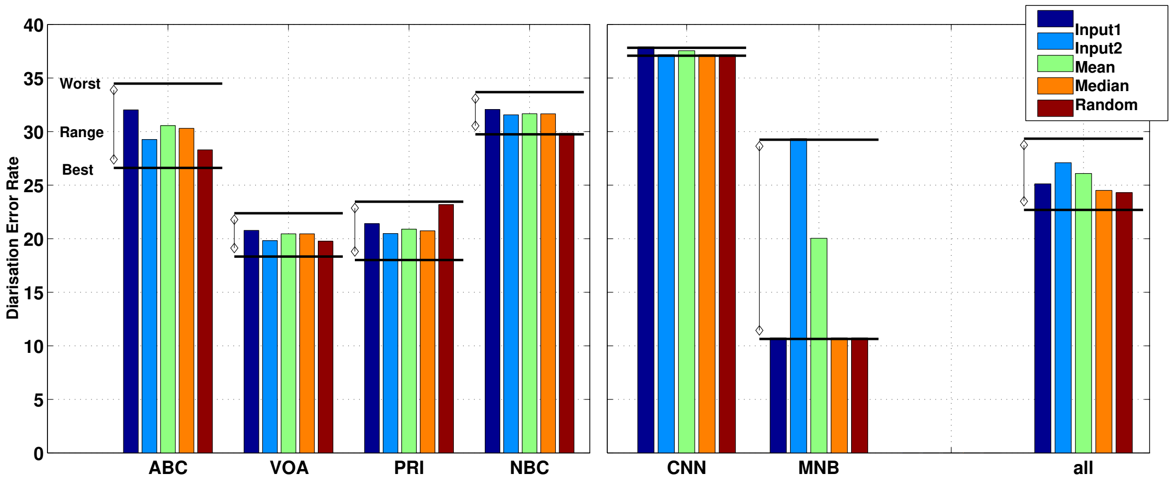

Figure 3: *Example 2: Graphical representation of the key DERs for the possible cluster voting outputs when 'misses' are not allowed.*

When misses are not allowed the following observations can be made from the results.

- The two inputs give DERs of 25.12% and 27.09%. The range of possible DERs when using the cluster voting output is increased by approx ± 2.34% to [22.79% → 29.44%]. This is a much greater increase than the ± 0.17% in Example 1 where the two inputs were much more similar.
- The number of possible outputs has dramatically increased. The sum over all 6 shows now being almost 25,000, as compared to 14 for Example 1.
- When the mean score for each show is taken separately this leads to an equivalent score of 26.09% over the 6-show set. This is also the mean of the two inputs.
- When taking the (lower) median for each show, the overall DER is 24.51%. This is less than both the inputs, showing that at least half of the combinations of the two inputs improve performance.
- When a possible output was selected at random for each show, the overall DER was 24.30%. This is fairly close to the median case and is also an improvement over both inputs.

- A single decision as to whether to split a supergroup of 473.6s duration into two speakers or keep it as a single speaker alters the DER for the MNB show from 10.74% to 29.34% and since the dataset only contains 6 shows in total this single decision can affect the overall DER on the bneval03 by around 3% absolute. This is an unfortunate consequence of having small data sets and a time-weighted scoring metric.
- Choosing all the outputs to be the same (25.19%) or all different (27.02%) gave DERs between the two inputs. This does not support the idea that a simple generic decision can be taken for all cases of conflict. A modified decision which considers other factors such as the size of the supergroup or the number of possibilities may be helpful, but how to reliably learn such rules with such a small amount of data is a very difficult problem.

6.3.1 Forcing Behaviour for Small Supergroups

An experiment was conducted which assigned a single unique speaker id for any supergroup whose total duration was less than a certain cut-off time. Since the diarisation error rate is time-weighted small supergroups generally have little effect on the final scores but add to the complexity of the problem and size of the CVOS. The results are given in Table 21 in Appendix B and illustrated graphically in Figure 4.

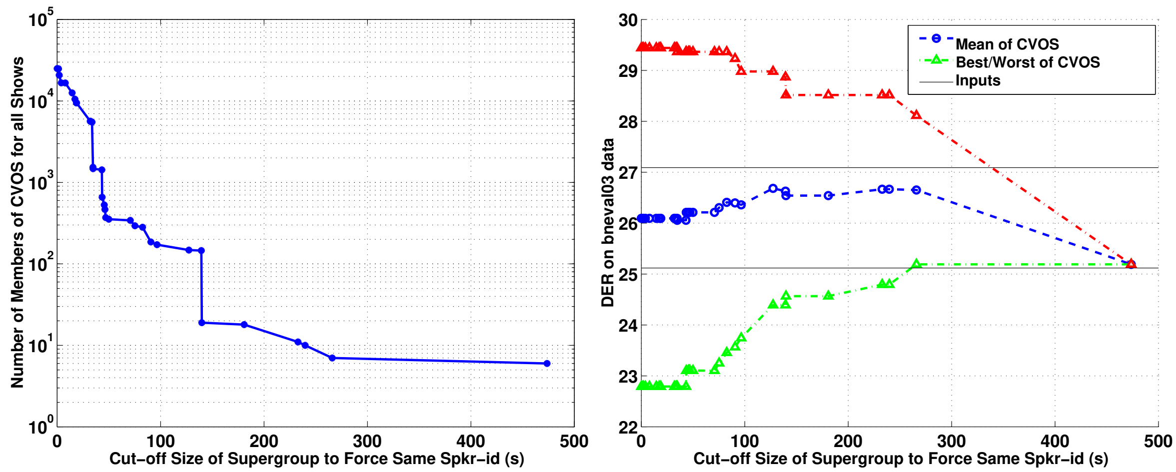

Figure 4: *Example 2: Graphical representation of the size of CVOS and range of DERs from the CVOS when forcing small supergroups to have a single unique speaker id. ('Misses' are not allowed.)*

As the small supergroup cut-off time threshold is increased the number of members of the CVOS decreases dramatically whilst the range of DERs producible from the CVOS decreases correspondingly. However, the best possible DER from the CVOS stays $\leq$ 23.1% for times up to around 70s, the corresponding drop in number of CVOS members being from 24,992 to 342.

This experiment shows that if a good method of picking the final output from the CVOS is in place it may speed up both the CVOS generation and judging processes without loss of performance by forcing small supergroups (of total duration under say 1 minute) to contain a single unique speaker.

6.4 Scoring All Possibilities from the CVOS when Allowing Misses for Conflicting Segments

The experiment described in section 6.3 was repeated allowing the supergroups to be missed out from the final output. The results when any combination of supergroup were allowed to be missing from the output are given in Table 12 and summarised graphically in Figure 5.

Condition		Show						
		ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
Input 1		32.03	20.78	21.40	32.06	37.92	10.74	25.12
Input 2		29.26	19.82	20.48	31.56	37.18	29.34	27.09
misses allowed	Num Poss	225	52245	1125	61965	66825	135	(182,520)
	Best	26.71	18.31	17.84	29.11	36.12	9.55	22.26
	Worst	41.46	25.99	34.44	39.74	49.31	48.03	38.81
	Mean	32.58	21.26	23.78	32.49	40.90	28.89	(29.13)
	Median	32.03	21.24	22.51	32.14	37.92	29.33	28.37
	Random	*29.25*	19.95	21.21	33.21	49.28	46.85	32.05
	All missed	41.46	25.99	34.18	39.02	48.25	46.85	38.31

Table 12: *Example 2: Summary of DERs for the possible cluster voting outputs when supergroup 'misses' are allowed.*

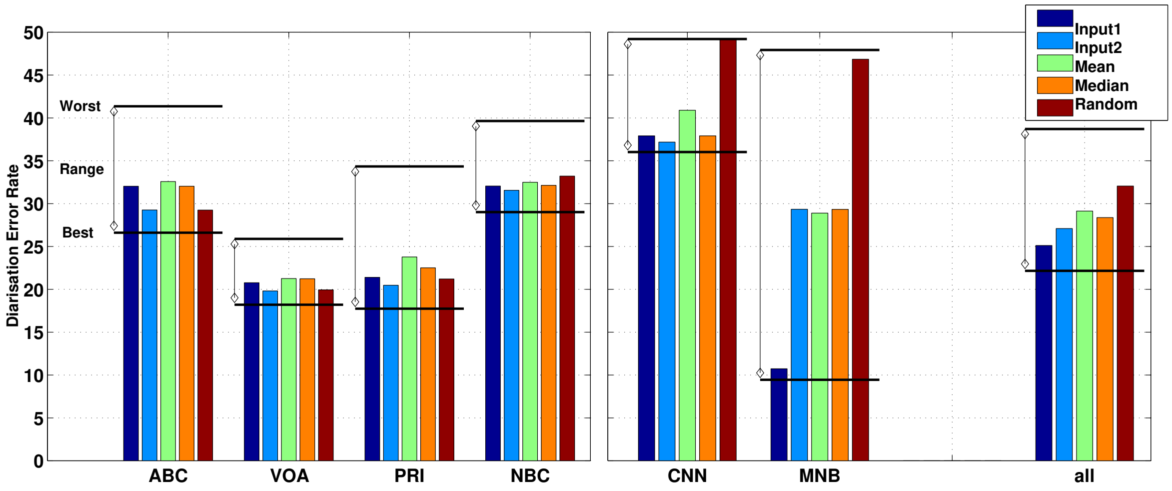

Figure 5: *Example 2: Graphical representation of the key DERs for the possible cluster voting outputs when 'misses' are allowed.*

If missing supergroups of conflicting segments out of the output *is allowed* the following observations can be made from the results.

- The best possible DER (22.26%) is slightly better than for the no-miss case (22.79%), however the worst DER drops from 29.44% to 38.81%, the mean drops from 26.09% to 29.13% and the median drops from 24.51% to 28.37% showing that in general allowing misses makes the performance considerably worse.
- The number of possible outputs from the cluster voting is increased by a factor of 7.3 over the no-miss case, giving 182,520 possibilities.

The increase in number of possible outputs and the decrease in general standard of these outputs when allowing conflicting segments to be omitted for Example 2 suggests that this is not a good strategy when the difference between the inputs is not very small. Therefore further work will concentrate on the case where no misses are allowed when generating the CVOS, although it is noted that it is possible to get small improvements in some cases by allowing misses in the CVOS.

6.5 Judging the CVOS Possibilities Directly

It is encouraging to note that there is a potential for a 2.34% absolute reduction in DER over the best input using the cluster-voting scheme for Example 2, and that both the median and a randomly chosen output from the list of possibilities (when misses are not allowed) give better performance than either of the inputs. However, a more theoretically sound method of deciding which member of the CVOS to use as the final output must be sought. Two BIC-based strategies as discussed in section 4.2 were implemented and tested using Example 2.

6.5.1 Standard BIC

A single full-covariance Gaussian model was built on the static-only PLP coefficients for each *cluster* in each supergroup. The weighted likelihood score given in equation 1:

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

was then calculated for each supergroup independently and the model set which gave the maximum BIC value for the specified α was selected for the final output. This was repeated for different values of the α parameter, including the value which was found to optimal for the BIC-based clustering on the development data, namely 7.25 (Tranter and Reynolds 2004). A complete set of results is given in Table 22 in Appendix B. A summary of the results are given in Table 13 and illustrated in Figure 6.

α	ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
0	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
3	*27.66*	19.94	*20.36*	***29.84***	37.92	29.34	26.72
6	30.30	19.94	*19.15*	32.06	37.92	**10.74**	*24.26*
7.25	32.85	19.94	*19.15*	32.06	37.92	**10.74**	*24.62*
9	33.63	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.85*
12	33.63	20.78	*19.15*	32.06	**37.18**	**10.74**	*24.79*
20	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
Input 1	32.03	20.78	21.40	32.06	37.92	**10.74**	25.12
Input 2	29.26	19.82	20.48	31.56	**37.18**	29.34	27.09
Best in CVOS	26.71	18.43	18.11	29.84	37.18	10.74	22.79
Worst in CVOS	34.58	22.48	23.56	33.78	37.92	29.34	29.44

Table 13: *Example 2: DERs when using the standard BIC formula to choose the 'best' alternative from the cluster voting output set. Numbers in italics are when the final output is better than either input. Numbers in bold match the best possible number from the CVOS. Each cluster is modelled using a full covariance Gaussian.*

The results show that the final output from the cluster-voting and BIC-judging system gives a lower DER than either of the inputs for a range in α values from 6 to 15, and that the lowest DER is 24.26% when α is 6, a 0.86% and 2.83% absolute improvement over the two inputs respectively.

The experiment was repeated using a Gaussian Mixture Model (GMM) built with diagonal covariances for each *cluster* in each supergroup instead of the previous full-covariance Gaussian. The number of free parameters per cluster in this case for a G-mixture GMM becomes:

$$\#M_{\text{GMM}} = 2dG + (G-1)$$

instead of the previous full-covariance single Gaussian which had:

$$\#M_{\text{fullcov}} = d + \frac{d(d+1)}{2}$$

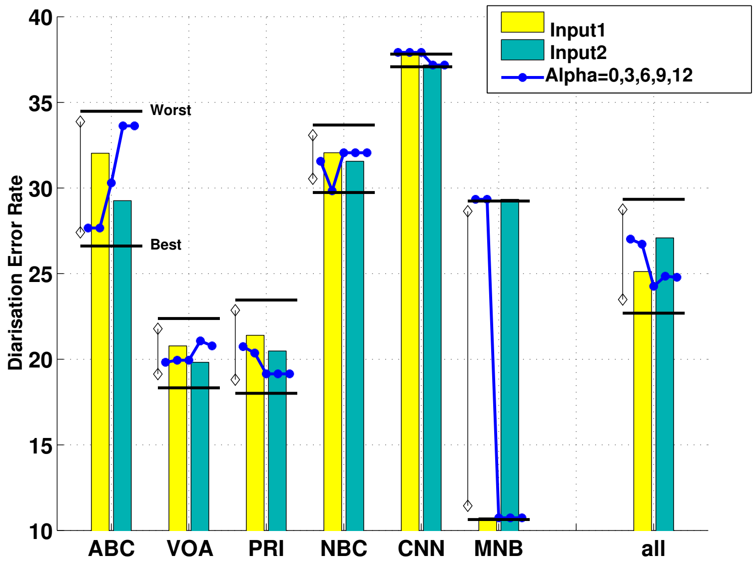

Figure 6: *Example 2: Graphical representation of the DERs when using the standard BIC formula to choose the 'best' alternative from the cluster voting output set. Each cluster is modelled using a full covariance Gaussian.*

The detailed results for a single diagonal covariance and {5,8,16,32,64,128} mixture diagonal covariance GMMs are given in Tables 22, 23 and 24 in Appendix B. The overall results on the 6-show bneval03 set are summarised in Table 14 and illustrated graphically in Figure 7. The values of α which give the best and worst DERs are given along with the range of α which improves (or worsens) performance. For the 16 or more mixture GMMs the value of α was varied by 0.1 for each step, whereas the ones with fewer parameters had α varied by 1.0 for each step.

Resegment Model	Best DER (α)	Worst DER (α)	DER (α=7.25)	DER < inputs α-range	DER between inputs α-range	DER > inputs α-range
full-cov	24.26 (6)	27.02 (0)	24.62	6 $\rightarrow$15	0 $\rightarrow$5, $\geq$ 16	-
1-mix diag-cov	23.90 (10)	27.02 (0)	26.72	10 $\rightarrow$20	0 $\rightarrow$9, 100	-
5-mix diag-cov	24.16 (5)	27.08 (4)	24.83	5 $\rightarrow$13	0 $\rightarrow$4, $\geq$ 14	-
8-mix diag-cov	24.63 (4)	27.28 (3)	24.85	4 $\rightarrow$7.25	0 $\rightarrow$2, $\geq$ 8	3
16-mix diag-cov	24.52 (2.5)	27.89 (2.1)	25.19	2.2 $\rightarrow$5.5	0 $\rightarrow$1.9, $\geq$ 5.6	2.0-2.1
32-mix diag-cov	24.23 (1.4)	27.02 (0)	25.19	1.2 $\rightarrow$2.4	0 $\rightarrow$1.1, $\geq$ 2.5	-
64-mix diag-cov	23.90 (0.7)	27.02 (0)	25.19	0.7 $\rightarrow$1.2	0 $\rightarrow$0.6, $\geq$ 1.3	-
128-mix diag-cov	23.76 (0.5)	27.02 (0)	25.19	0.5 $\rightarrow$0.7	0 $\rightarrow$0.4, $\geq$ 0.8	-

Table 14: *Example 2: Summary of the results when using the standard BIC formula to choose the 'best' alternative from the CVOS. The inputs give DERs of 25.12 and 27.09% respectively.*

The results show that for every case it is possible to find a range of α values where the chosen output is better than both the inputs. Both the range of α which reduces the DER, and the absolute value of the optimal α parameter decreases as the number of parameters increase. The variation of the optimal DER performance and corresponding α value with the number of parameters is given in Figure 8. Initially the best DER gets worse as the number of parameters increases, but this slowly recovers until it matches the case of a single diagonal covariance with that of a 64-mixture GMM. Since the range

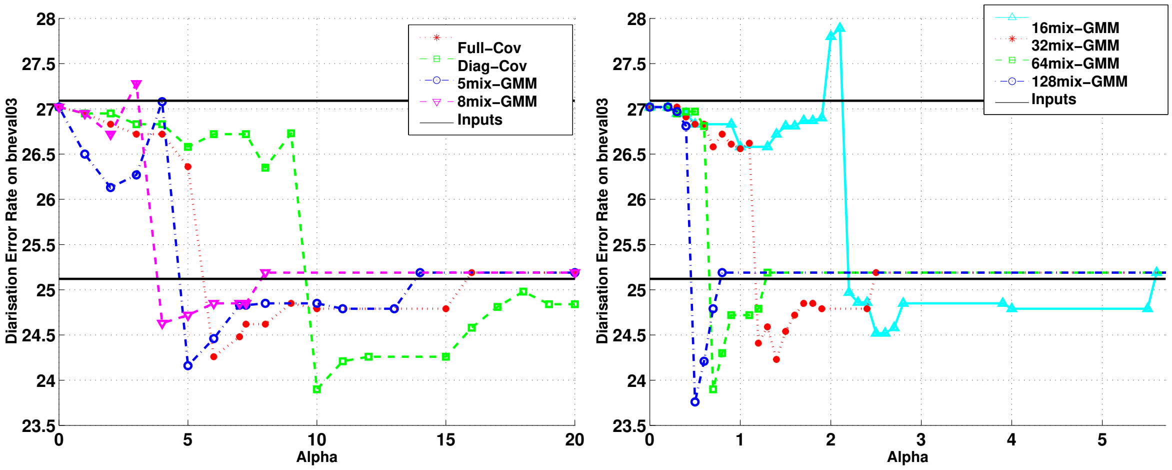

Figure 7: *Example 2: Graphical representation of the DERs when using the standard BIC formula to choose the 'best' alternative from the cluster voting output set when modelling each cluster with a diagonal-covariance GMM or a full-covariance Gaussian. The sudden improvement in performance in each case occurs when the single critical decision for the MNB show is made successfully.*

of α which improves the DER is much greater for the former case, and only slightly improvements are gained from increasing the number of mixtures in the GMM further, it appears the wisest choice for this experiment may be to take the single diagonal covariance representation for each cluster. This gives a DER of 23.90% which is an improvement of 1.22% and 3.19% absolute over the two inputs respectively. The optimal performance in these experiments (for a 128-mixture GMM) was 23.76% DER, an improvement of 1.36% and 3.33% absolute over the two inputs.

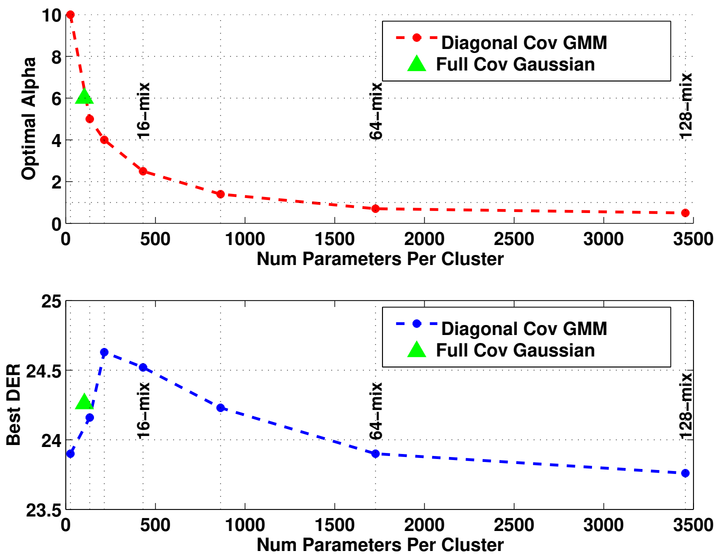

Figure 8: *Example 2: Graphical representation of the best DER when using the standard BIC formula to choose the 'best' alternative from the cluster voting output set when modelling each cluster with a diagonal-covariance GMM or a full-covariance Gaussian. The optimal α for each case is also given.The inputs give DERs of 25.12% and 27.09% respectively.*

6.5.2 Equal-Parameter BIC

An alternative interpretation of the BIC-based judging scheme, discussed in section 4.2.2, where the need for the penalty term is eliminated by using the same number of parameters in the alternative model-sets was implemented. Each supergroup was treated independently and the number of parameters for each cluster(=speaker-id) was made proportional to the number of resegments for that speaker. For example, if a supergroup consists of two resegments which can either be labelled as the same speaker or as different speakers, the model set for the different speakers consisted of one base-model for each speaker, whereas the combined model consisted of a mixture containing the equivalent of two base-models, trained on all of the data. Different parameterisations for the base-model were investigated and the results are given in Table 15 and illustrated in Figure 9.

Base Model	ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
1 full-Gaussian (s)	32.85	20.48	21.15	32.06	**37.18**	**10.74**	*25.02*
1 diag-covariance (s+d)	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
5-mix GMM diag cov (s+d)	32.85	21.17	21.15	32.06	**37.18**	**10.74**	25.16
1 diag-covariance (s)	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
5-mix GMM diag cov (s)	31.89	20.80	21.15	*31.05*	**37.18**	**10.74**	*24.80*
10-mix GMM diag cov (s)	30.30	19.94	***18.11***	*31.05*	**37.18**	**10.74**	*23.79*
15-mix GMM diag cov (s)	30.30	*19.27*	***18.11***	***29.84***	**37.18**	**10.74**	*23.48*
20-mix GMM diag cov (s)	30.30	*19.27*	***18.11***	*30.18*	**37.18**	**10.74**	*23.53*
25-mix GMM diag cov (s)	30.30	*19.27*	***18.11***	*30.50*	**37.18**	**10.74**	*23.57*
32-mix GMM diag cov (s)	32.84	*19.27*	*20.10*	*30.39*	37.92	**10.74**	*24.42*
Input 1	32.03	20.78	21.40	32.06	37.92	**10.74**	25.12
Input 2	29.26	19.82	20.48	31.56	**37.18**	29.34	27.09
Best Possible from CVOS	26.71	18.43	18.11	29.84	37.18	10.74	22.79
Worst Possible from CVOS	34.58	22.48	23.56	33.78	37.92	29.34	29.44

Table 15: *Example 2: DERs when using the equal-parameter BIC method to choose the 'best' alternative from the cluster voting output set. Numbers in italics are when the final output is better than either input. Numbers in bold match the best possible number from the CVOS. (s) means static coefficients only, whereas (s+d) means statics and deltas.*

The results show that almost every single representation gives an improvement over the two inputs using this method. When the base model is a full-covariance Gaussian the DER is 25.02% a very small improvement over the best input (25.12%). Using a single diagonal covariance Gaussian gives a slight degradation in performance (25.19%) but the performance improves dramatically as the number of mixtures in the GMM is increased. Adding in delta information into the feature vector did not seem to help performance at all.

The optimal performance occurs when using a 15-mixture GMM for the base-model. This gives a DER of 23.48% which is an absolute improvement of 1.64% over the best input and is only 0.69% off the best possible choice from the CVOS. In fact for 4 of the 6 shows (PRI/NBC/CNN/MNB) this case does indeed choose the best possible option from the CVOS.

Note that no knowledge of the number of frames in the resegments or supergroups was used in this experiment, and it may be possible to get further improvements if the number of Gaussians in the GMMs was directly related to the number of frames in the speaker-cluster being considered. In this way the total number of parameters for competing model-sets would still remain constant, but more parameters would be used to model speakers with more data associated with them. This is effectively what happens if this type of scheme is used for speaker clustering and the initial segmentation is uniform (i.e. the segments all start off with the same number of frames and the same number of parameters modelling them as is done in (Ajmera and Wooters 2003).)

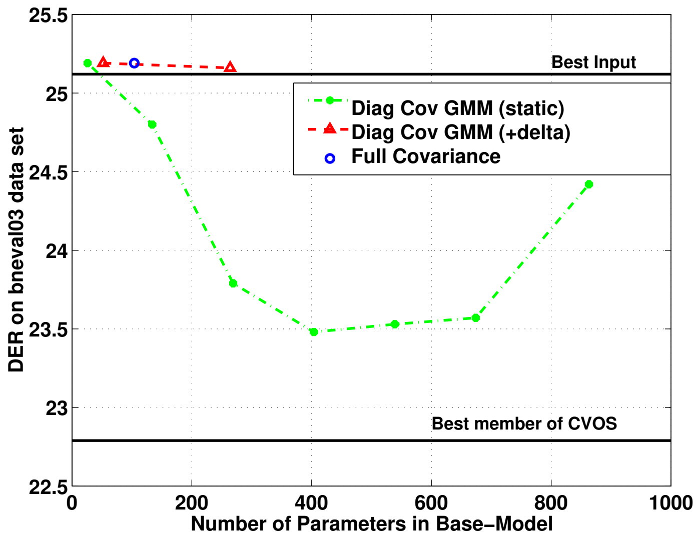

Figure 9: *Example 2: Graphical representation of DERs when using the equal-parameter BIC method to choose the 'best' alternative from the cluster voting output set. The inputs give DERs of 25.12% and 27.09% respectively.*

6.6 Summary of Experimental Results

The key results from Example 2 are summarised in Table 16 and illustrated in Figure 10.

System	ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
Input 1	32.03	20.78	21.40	32.06	37.92	**10.74**	25.12
Input 2	29.26	19.82	20.48	31.56	**37.18**	29.34	27.09
Best Possible from CVOS	26.71	18.43	18.11	29.84	37.18	10.74	22.79
Worst Possible from CVOS	34.58	22.48	23.56	33.78	37.92	29.34	29.44
Random	*28.30*	*19.76*	23.18	***29.84***	**37.18**	**10.74**	*24.30*
All the same	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
All different	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
†Standard BIC (full-cov)	30.30	19.94	*19.15*	32.06	37.92	**10.74**	*24.26*
†Standard BIC (diag-cov)	30.30	19.94	***18.11***	*31.05*	37.92	**10.74**	*23.90*
†Standard BIC (128mix GMM)	*27.66*	20.78	*19.15*	*30.83*	**37.18**	**10.74**	*23.76*
Equal-param BIC (full-cov)	32.85	20.48	21.15	32.06	**37.18**	**10.74**	*25.02*
Equal-param BIC (diag-cov)	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
Equal-param BIC (15mix GMM)	30.30	*19.27*	***18.11***	***29.84***	**37.18**	**10.74**	*23.48*

Table 16: *Example 2 : Summary of Key DERs. Numbers in italics are when the final output is better than either input. Numbers in bold match the best possible number from the CVOS. † Results are for the optimal α value in the standard BIC formulation.*

The cluster voting scheme gives a set of around 25,000 possible output clusterings, the cluster voting output set (CVOS) which range in DER from 29.44% to 22.79%, as compared to the two inputs with DER of 25.12% and 27.09%.

Several methods of choosing a member of the CVOS as the final output have been investigated. The (lower) median and picking a member at random both improved the performance over both inputs, giving DERs of 24.51% and 24.30% respectively.

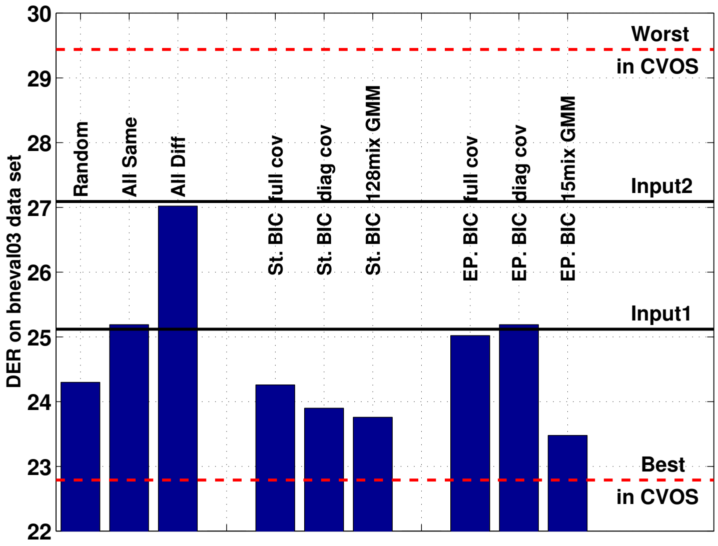

Figure 10: *Example 2 : Graphical Representation of the key DERs given in Table 16.*

A conventional BIC-based method to select the best model-set from the CVOS gave further improvements in performance, the case of a single full-covariance Gaussian for each clustering giving a DER of 24.26% for $\alpha = 6$ and an improvement over both inputs for $6 \leq \alpha \leq 15$. Using a single diagonal covariance Gaussian gave a DER of 23.90% for $\alpha = 10$ and improvement over both inputs for $10 \leq \alpha \leq 20$. The best score from this method was 23.76% for $\alpha = 0.5$ when using a 128-mixture GMM, however the range of α which improved the DER over both inputs in this case was only $0.5 \leq \alpha \leq 0.7$.

A second judging scheme based on an equal-parameter implementation of the BIC formulation was also implemented. This gave only small improvements when using a full covariance Gaussian as the base-model (DER=25.02%) but much greater improvements when using diagonal covariance GMMs instead - the DER for this case being under 24% for base-models with {10,15,20,25} mixtures in the base-model GMM. The best performance in this case occurred when using a 15-mixture GMM for the base-model, giving a DER of 23.48%, an absolute improvement of 1.64% over the best input.

Recent work in speaker diarisation at MIT Lincoln Labs (Reynolds and Torres-Carrasquillo 2004) and LIMSI (Gauvain, Lamel, Schwenk, Adda, Barras, Chen, Lefevre, Meignier and Messaudi 2004) has confirmed the experience of CUED that better results can be obtained if segments are modelled using a single full-covariance Gaussian rather than a GMM using diagonal covariance Gaussians. However, this was not found to be the case for either the conventional or equal-parameter BIC judging schemes in this experiment, where the best DER was reduced by a further 0.5-1.5% absolute when switching from a single full covariance to diagonal covariance GMM-based implementation.

7 Conclusions

A cluster voting scheme has been defined which takes the output of different speaker diarisation systems as its inputs and tries to produce a new speaker clustering which gives a lower diarisation error rate (DER). This paper has described this system in detail for the case of two inputs, but the framework can be easily extended theoretically to work with more inputs if required.

The process is split into two main parts. The first part involves producing a set of possible clusterings which maximise the cluster voting metric (CVM) to produce the cluster voting output set (CVOS), whilst the second part defines a way of choosing the final output from the CVOS. For this work the CVM was defined as the sum of the overlap in time between the output clustering and the two inputs under the one-to-one speaker mapping rules used in the definition of the DER. This is equivalent to minimising the sum of the DER from the output to the two inputs.

Due to the nature of this mapping, there are several different clusterings which maximise the CVM - including both the inputs. If there are more than two inputs it may be possible to reduce the number of members of the CVOS using confidence scores to weight the inputs, but this is not applicable for the two-input case, since the highest weighted input would simply become the output.

Generating all possible output clusterings in order to find all the possible CVM scores is computationally complex, with order approximately N! where N is the number of segments. Several methods of reducing the complexity by resegmenting the data, fast-tracking segments which do not conflict to the output, and effectively 'factorising' the segments into independent supergroups have been discussed and shown to allow a practical implementation of the method on some real data. Further reduction of the number of possible output clusterings which need to be generated in order to find those which maximise the CVM can be made by using the information from the one-to-one mapping between the inputs to restrict the possibilities.

It has also been shown that it is possible to reduce the complexity further without unduly affecting the final score. For example, this can be accomplished by assigning all the segments in a supergroup to the same unique speaker-id if the duration of the supergroup is small. This means that not all members of the theoretical CVOS will be generated, but since the overall diarisation scoring is time-weighted, getting the speaker-id of small segments correct is not generally critical and thus the final score will probably not be unduly affected despite this simplification. Similar rules which automatically assign certain speaker-ids or even miss some segments out from the final output completely can easily be incorporated to attempt to reduce the size of the CVOS without unduly affecting the performance of the final output.

Once the CVOS has been generated, a method of picking the final output from it must be defined. This paper has described several methods of doing this, ranging from simple rule-based strategies such as forcing the segments in a supergroup to either have the same or all different speaker-ids, or more conventional model-selection techniques based on the Bayes Information Criterion (BIC).

Results have been presented on two experiments on the RT-03s Broadcast News evaluation data. The first as a proof-of-concept using two diarisation systems as input which were identical apart from a slight difference in final stopping criterion, whilst the second used the best two CUED diarisation systems available in December 2003. The results showed that it was theoretically possible to reduce the DER from the 25.12/27.09% of these two systems to 22.79% using this method, and the two BIC-based judging schemes discussed in this paper produced DERs of 23.76% and 23.48% - providing a 1.64% absolute reduction in DER over the previous best CUED system.

A The Bell Series

The Bell series (Sloane's series A000110 (Sloane 2003)) is named after Eric Temple Bell (1883-1960) who did some early work on this series(Bell 1934). It starts

```
[ 1 1 2 5 15 52 203 877 4140 21147 115975 ...]
```

and represents the number of ways a set with n elements can be partitioned into disjoint, non-empty sets. For example the set {A,B,C} can be partioned 5 ways into

- {{A}, {B}, {C}} = 1 2 3
- {{A,B}, {C}} = 1 1 2
- {{A,C}, {B}} = 1 2 1
- {{A}, {B,C}} = 1 2 2
- {{A,B,C}} = 1 1 1

The nth Bell number can be computed using the formula

$$B(n) = \sum_{k=0}^{n} S_n^{(k)}$$

where $S_n^{(k)}$ represents the Stirling numbers of the second kind - i.e. the number of ways a set with n elements can be partioned into k disjoint, non-empty subsets. These numbers can be computed recursively using the formula:

$$S_n^(k) = S_{n-1}^{(k-1)} + kS_{n-1}^{(k)}$$

The Bell numbers also give the coefficients of the Maclauren expansion of

$$e^{e^x} = e(1 + \frac{1x}{1!} + \frac{2x^2}{2!} + \frac{5x^3}{3!} + ..$$

and thus have a generating function of

$$e^{(e^x-1)} = \sum_{n\geq 0} \frac{B_n x^n}{n!}$$

They can be generated recursively using

$$B_{n+1} = \sum_{i=0}^{n} B_i C(n,i); \qquad C(n,i) = \frac{n!}{i!(n-i)!}$$

and can be shown using a Bell Triangle. Start with a row with the number one, each new row begins with the last number of the previous row and continues to the right adding each number to the number above it to get the next number in the row. The Bell numbers appear down the left hand side:

1						Start with one
1	2					Start with one add 1+1 to get 2
2	3	5				Start with two, add 2+1=3, 3+2=5
5	7	10	15			Start with five, add 5+2=7, 7+3=10, 10+5=15..etc..
15	20	27	37	52		
52	67	87	114	151	203	

Information about Bell number sequence was obtained from (Bath University WWW site 2004) and (Weisstein 2004).

B Comprehensive Results

B.1 Example 1 : Cluster Voting on Two Very Similar Systems

This appendix gives the results from the experiments in section 5.

VOA								CNN			NBC	
g0	g1	g2	DER	g0	g1	g2	DER	g0	g1	DER	g0	DER
d	d	d	19.94	m	d	m	22.43	m	m	36.46	m	31.74
s	d	d	20.34	s	s	m	22.52	m	d	36.86	[s|d]	32.06 *i1,*i2
d	s	d	20.76	m	s	s	22.75	m	s	36.90		
d	d	s	20.78 *i1	m	s	m	23.26	[s|d]	m	37.52		
m	d	d	21.08	d	m	d	29.46	[s|d]	d	37.92 *i1		
s	s	d	21.17 *i2	s	m	d	29.86	[s|d]	s	37.96 *i2		
s	d	s	21.18	d	m	s	30.30					
d	d	m	21.29	m	m	d	30.60					
d	s	s	21.60	s	m	s	30.70					
s	d	m	21.69	d	m	m	30.81					
m	s	d	21.91	s	m	m	31.22					
m	d	s	21.92	m	m	s	31.44					
s	s	s	22.01	m	m	m	31.96					
d	s	m	22.12									

Table 17: *Example 1: Diarisation Error Rate (DER) for all possible outputs from the CVOS (including missing conflicts out). DERs are generated against the true reference data. (s) means the cluster group has the same cluster-id for the 2 resegments, (d) means they are different, and (m) means they are missed out from the output completely. *ix is the same as input x.*

Condition		Show			
		VOA	CNN	NBC	TOTAL (all 6)
Input 1		20.78	37.92	32.06	25.12
Input 2		21.17	37.96	32.06	25.21
No miss allowed	Num Poss	8	4	2	(14)
	Best	19.94	37.92	32.06	24.95
	Worst	22.01	37.96	32.06	25.38
	Mean	20.97	37.94	**32.06**	25.17
	Median*	20.78	**37.92**	**32.06**	25.12
	Random	20.78	**37.92**	**32.06**	25.12
	All the same	22.01	37.96	**32.06**	25.38
	All different	***19.94***	**37.92**	**32.06**	***24.95***
miss allowed	Num Poss	27	9	3	(39)
	Best	19.94	36.46	31.74	24.68
	Worst	31.96	37.96	32.06	27.44
	Mean	24.63	*37.45*	*31.95*	25.83
	Median	22.12	*37.52*	32.06	25.34
	Random	21.21	37.92	*31.74*	25.16
	All missed	31.96	**36.46**	**31.74**	27.17

Table 18: *Example 1: Summary of DERs for the possible cluster voting outputs. [* The lower median is taken here i.e. the n/2 th entry in the list of n possibilities]*

B.2 Example 2 : Cluster Voting on CUED's Two Best Systems

This appendix gives the results from all the experiments in section 6. In some cases more detail is provided than is in the main paper.

Condition		ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
Input 1		32.03	20.78	21.40	32.06	37.92	10.74	25.12
Input 2		29.26	19.82	20.48	31.56	37.18	29.34	27.09
No misses allowed	Num Poss	128	8192	256	8192	8192	32	(24,992)
	Best	26.71	18.43	18.11	29.84	37.18	10.74	22.79
	Worst	34.58	22.48	23.56	33.78	37.92	29.34	29.44
	Mean	30.56	20.45	20.89	31.67	37.55	20.04	(26.09)
	Median*	30.30	20.45	20.74	31.66	**37.18**	**10.74**	*24.51*
	Random	*28.30*	*19.76*	23.18	***29.84***	**37.18**	**10.74**	*24.30*
	All the same	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
	All different	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02

Table 19: *Example 2: Summary of DERs for the possible cluster voting outputs when 'misses' are not allowed. [* The lower median is taken here i.e. the n/2 th entry in the list of n possibilities.]*

Condition		ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
Input 1		32.03	20.78	21.40	32.06	37.92	10.74	25.12
Input 2		29.26	19.82	20.48	31.56	37.18	29.34	27.09
misses allowed	Num Poss	225	52245	1125	61965	66825	135	(182,520)
	Best	26.71	18.31	17.84	29.11	36.12	9.55	22.26
	Worst	41.46	25.99	34.44	39.74	49.31	48.03	38.81
	Mean	32.58	21.26	23.78	32.49	40.90	28.89	(29.13)
	Median	32.03	21.24	22.51	32.14	37.92	29.33	28.37
	Random	*29.25*	19.95	21.21	33.21	49.28	46.85	32.05
	All missed	41.46	25.99	34.18	39.02	48.25	46.85	38.31

Table 20: *Example 2: Summary of DERs for the possible cluster voting outputs when supergroup 'misses' are allowed.*

Show	Supergroup cut-off(s)	Size of CVOS Time	Num Poss	Diarisation Error Rate Mean	Median	Best	Worst
ABC	0	3912.28	128	30.56	30.30	26.71	34.58
ABC	90.56	975.34	32	30.48	30.04	27.48	33.63
ABC	127.3	262.00	8	32.75	32.67	32.03	33.63
ABC	232.87	33.63	1	33.63	33.63	33.63	33.63
VOA	0	167514.24	8192	20.45	20.45	18.43	22.48
VOA	14.46	83757.12	4096	20.45	20.45	18.43	22.48
VOA	17.14	41878.56	2048	20.45	20.45	18.43	22.48
VOA	18.6	20939.28	1024	20.45	20.45	18.43	22.48
VOA	43.46	5427.96	256	21.20	21.17	19.94	22.48
VOA	45.52	2713.98	128	21.20	21.17	19.94	22.48
VOA	139.85	20.78	1	20.78	20.78	20.78	20.78
PRI	0	5346.88	256	20.89	20.74	18.11	23.56
PRI	1.28	2673.44	128	20.89	20.74	18.11	23.56
PRI	34.66	1324.72	64	20.70	20.36	18.11	23.18
PRI	43.19	331.18	16	20.70	20.36	18.11	23.18
PRI	82.51	84.88	4	21.22	21.15	19.15	23.18
PRI	265.98	21.15	1	21.15	21.15	21.15	21.15
NBC	0	259432.96	8192	31.67	31.66	29.84	33.78
NBC	1.98	129716.48	4096	31.67	31.66	29.84	33.78
NBC	31.97	8107.28	256	31.67	31.66	29.84	33.78
NBC	33.72	4053.64	128	31.67	31.66	29.84	33.78
NBC	46.2	2026.82	64	31.67	31.66	29.84	33.78
NBC	75.22	517.02	16	32.31	32.06	30.84	33.78
NBC	96.56	64.12	2	32.06	32.06	32.06	32.06
NBC	180.75	32.06	1	32.06	32.06	32.06	32.06
CNN	0	307609.60	8192	37.55	37.18	37.18	37.92
CNN	3.89	153804.80	4096	37.55	37.18	37.18	37.92
CNN	34.62	4806.40	128	37.55	37.18	37.18	37.92
CNN	46.9	1201.60	32	37.55	37.18	37.18	37.92
CNN	49.92	600.80	16	37.55	37.18	37.18	37.92
CNN	70.69	150.20	4	37.55	37.18	37.18	37.92
CNN	139.37	74.36	2	37.18	37.18	37.18	37.18
CNN	239.75	37.18	1	37.18	37.18	37.18	37.18
MNB	0	641.28	32	20.04	10.74	10.74	29.34
MNB	7.81	320.64	16	20.04	10.74	10.74	29.34
MNB	18.73	80.16	4	20.04	10.74	10.74	29.34
MNB	32.08	40.08	2	20.04	10.74	10.74	29.34
MNB	473.61	10.74	1	10.74	10.74	10.74	10.74

Table 21: *Example 2: Size of CVOS and range of DERs from the CVOS when forcing small supergroups to have a single unique speaker id. ('Misses' are not allowed.)*

Covariance	α	ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
full	0	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
full	1	*27.66*	19.82	*20.36*	31.56	37.92	29.34	26.95
full	2	*27.66*	*19.27*	*20.36*	31.56	37.92	29.34	26.83
full	3	*27.66*	19.94	*20.36*	***29.84***	37.92	29.34	26.72
full	4	***26.71***	19.94	*20.36*	*30.74*	37.92	29.34	26.72
full	5	***26.71***	19.94	***18.11***	*31.40*	37.92	29.34	26.36
full	6	30.30	19.94	*19.15*	32.06	37.92	**10.74**	*24.26*
full	7	31.89	19.94	*19.15*	32.06	37.92	**10.74**	*24.48*
full	7.25	32.85	19.94	*19.15*	32.06	37.92	**10.74**	*24.62*
full	8	32.85	19.94	*19.15*	32.06	37.92	**10.74**	*24.62*
full	9	33.63	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.85*
full	10-15	33.63	20.78	*19.15*	32.06	**37.18**	**10.74**	*24.79*
full	16-100	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
diagonal	0	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
diagonal	1-2	*27.66*	19.82	*20.36*	31.56	37.92	29.34	26.95
diagonal	3-4	*27.66*	*19.27*	*20.36*	31.56	37.92	29.34	26.83
diagonal	5	*27.66*	*19.27*	*20.36*	***29.84***	37.92	29.34	26.58
diagonal	6-7	*27.66*	19.94	*20.36*	***29.84***	37.92	29.34	26.72
diagonal	7.25	*27.66*	19.94	*20.36*	***29.84***	37.92	29.34	26.72
diagonal	8	*27.66*	19.94	***18.11***	*30.39*	37.92	29.34	26.35
diagonal	9	30.30	19.94	***18.11***	*30.39*	37.92	29.34	26.73
diagonal	10	30.30	19.94	***18.11***	*31.05*	37.92	**10.74**	*23.90*
diagonal	11	30.30	19.94	*19.15*	31.71	37.92	**10.74**	*24.21*
diagonal	12-15	30.30	19.94	*19.15*	32.06	37.92	**10.74**	*24.26*
diagonal	16	32.03	20.33	*19.15*	32.06	37.92	**10.74**	*24.58*
diagonal	17	33.63	20.33	*19.15*	32.06	37.92	**10.74**	*24.81*
diagonal	18	33.63	21.17	*19.15*	32.06	37.92	**10.74**	*24.98*
diagonal	19-20	33.63	20.48	*19.15*	32.06	37.92	**10.74**	*24.84*
diagonal	100	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
Input 1		32.03	20.78	21.40	32.06	37.92	**10.74**	25.12
Input 2		29.26	19.82	20.48	31.56	**37.18**	29.34	27.09
Best in CVOS		26.71	18.43	18.11	29.84	37.18	10.74	22.79
Worst in CVOS		34.58	22.48	23.56	33.78	37.92	29.34	29.44

Table 22: *Example 2: DERs when using the standard BIC formula to choose the 'best' alternative from the cluster voting output set. Numbers in italics are when the final output is better than either input. Numbers in bold match the best possible number from the CVOS. Each cluster is modelled using a single full or diagonal covariance Gaussian.* α *was incremented in steps of 1.*

Num-Mix	α	ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
5	0	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
5	1	*27.66*	19.82	***18.11***	31.56	37.92	29.34	26.50
5	2	*27.66*	*19.27*	***18.11***	***29.84***	37.92	29.34	26.13
5	3	*27.66*	19.94	***18.11***	***29.84***	37.92	29.34	26.27
5	4	30.30	19.94	*19.15*	*31.40*	37.92	29.34	27.08
5	5	30.30	19.94	*19.15*	*31.40*	37.92	**10.74**	*24.16*
5	6	31.72	19.94	*19.15*	32.06	37.92	**10.74**	*24.46*
5	7/7.25	32.68	21.07	*19.15*	32.06	37.92	**10.74**	*24.83*
5	8-10	33.63	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.85*
5	11-13	33.63	20.78	*19.15*	32.06	**37.18**	**10.74**	*24.79*
5	14-100	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
8	0	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
8	1	*27.66*	19.82	*20.36*	31.56	37.92	29.34	26.95
8	2	*27.66*	19.94	*20.36*	***29.84***	37.92	29.34	26.72
8	3	31.25	20.78	***18.11***	32.06	37.92	29.34	27.28
8	4	31.72	20.78	*19.15*	32.06	37.92	**10.74**	*24.63*
8	5	32.68	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.72*
8	6-7/7.25	33.63	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.85*
8	8-20,100	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
16	0-0.2	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
16	0.3-0.4	*27.66*	19.82	*20.36*	31.56	37.92	29.34	26.95
16	0.5-0.9	*27.66*	*19.27*	*20.36*	31.56	37.92	29.34	26.83
16	1.0-1.3	*27.66*	*19.27*	*20.36*	***29.84***	37.92	29.34	26.58
16	1.4	*27.66*	19.94	*20.36*	***29.84***	37.92	29.34	26.72
16	1.5-1.6	*27.66*	19.94	*20.36*	*30.49*	37.92	29.34	26.81
16	1.7-1.8	*27.66*	19.94	*20.36*	*30.83*	37.92	29.34	26.87
16	1.9	***26.71***	20.78	*20.36*	*30.83*	37.92	29.34	26.90
16	2.0	30.30	20.78	21.40	32.06	37.92	29.34	27.80
16	2.1	30.94	20.78	21.40	32.06	37.92	29.34	27.89
16	2.2	30.94	20.78	21.40	32.06	37.92	10.74	*24.97*
16	2.3-2.4	30.94	20.78	21.40	32.06	**37.18**	10.74	*24.86*
16	2.5-2.6	31.72	20.78	*19.15*	32.06	**37.18**	10.74	*24.52*
16	2.7	31.72	21.07	*19.15*	32.06	**37.18**	10.74	*24.58*
16	2.8-3.9	33.63	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.85*
16	4.0-5.5	33.63	20.78	*19.15*	32.06	**37.18**	**10.74**	*24.79*
16	5.6-20,100	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
Input 1		32.03	20.78	21.40	32.06	37.92	**10.74**	25.12
Input 2		29.26	19.82	20.48	31.56	**37.18**	29.34	27.09
Best in CVOS		26.71	18.43	18.11	29.84	37.18	10.74	22.79
Worst in CVOS		34.58	22.48	23.56	33.78	37.92	29.34	29.44

Table 23: *Example 2: DERs when using the standard BIC formula to choose the 'best' alternative from the cluster voting output set. Numbers in italics are when the final output is better than either input. Numbers in bold match the best possible number from the CVOS. Each cluster is modelled using a diagonal covariance GMM. α was incremented in steps of 1 for 5 and 8 mixture components, and 0.1 for 16 mixture components.*

Num-Mix	α	ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
32	0-0.3	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
32	0.4	*27.66*	*19.27*	20.74	31.56	37.92	29.34	26.91
32	0.5-0.6	*27.66*	*19.27*	*20.36*	31.56	37.92	29.34	26.83
32	0.7	*27.66*	*19.27*	*20.36*	***29.84***	37.92	29.34	26.58
32	0.8	*27.66*	19.94	*20.36*	***29.84***	37.92	29.34	26.72
32	0.9	*27.66*	19.94	*20.36*	***29.84***	**37.18**	29.34	26.61
32	1.0	***26.71***	19.94	*20.36*	*30.39*	**37.18**	29.34	26.56
32	1.1	***26.71***	19.94	*20.36*	*30.83*	**37.18**	29.34	26.62
32	1.2	30.30	19.94	21.40	*30.83*	**37.18**	**10.74**	*24.41*
32	1.3	30.30	20.78	21.40	*30.83*	**37.18**	**10.74**	*24.59*
32	1.4	30.94	20.78	*19.15*	*30.83*	**37.18**	**10.74**	*24.23*
32	1.5	31.89	20.78	*19.15*	32.06	**37.18**	**10.74**	*24.54*
32	1.6	32.67	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.72*
32	1.7-1.8	33.63	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.85*
32	1.9-2.4	33.63	20.78	*19.15*	32.06	**37.18**	**10.74**	*24.79*
32	2.5-20,100	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
64	0-0.2	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
64	0.3	*27.66*	19.82	*20.36*	31.56	37.92	29.34	26.95
64	0.4-0.5	*27.66*	19.94	*20.36*	31.56	37.92	29.34	26.97
64	0.6	*27.66*	19.94	*20.36*	*30.49*	37.92	29.34	26.81
64	0.7	***26.71***	19.94	21.40	*30.83*	**37.18**	**10.74**	*23.90*
64	0.8	*28.30*	20.78	21.40	*30.83*	**37.18**	**10.74**	*24.30*
64	0.9-1.1	33.63	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.72*
64	1.2	33.63	20.78	*19.15*	32.06	**37.18**	**10.74**	*24.79*
64	1.3-20,100	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
128	0-0.2	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
128	0.3	*27.66*	19.94	*20.36*	31.56	37.92	29.34	26.97
128	0.4	*27.66*	19.94	*20.36*	*30.49*	37.92	29.34	26.81
128	0.5	*27.66*	20.78	*19.15*	*30.83*	**37.18**	**10.74**	*23.76*
128	0.6	*29.08*	21.07	*19.15*	32.06	**37.18**	**10.74**	*24.21*
128	0.7	33.63	20.78	*19.15*	32.06	**37.18**	**10.74**	*24.79*
128	0.8-20,100	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
Input 1		32.03	20.78	21.40	32.06	37.92	**10.74**	25.12
Input 2		29.26	19.82	20.48	31.56	**37.18**	29.34	27.09
Best in CVOS		26.71	18.43	18.11	29.84	37.18	10.74	22.79
Worst in CVOS		34.58	22.48	23.56	33.78	37.92	29.34	29.44

Table 24: *Example 2: DERs when using the standard BIC formula to choose the 'best' alternative from the cluster voting output set. Numbers in italics are when the final output is better than either input. Numbers in bold match the best possible number from the CVOS. Each cluster is modelled using a diagonal covariance GMM. α was incremented in steps of 0.1 .*

Resegment Model	Best DER (α)	Worst DER (α)	DER (α=7.25)	DER < inputs α-range	DER between inputs α-range	DER > inputs α-range
full-cov	24.26 (6)	27.02 (0)	24.62	6 $\rightarrow$15	0 $\rightarrow$5, $\geq$ 16	-
1-mix diag-cov	23.90 (10)	27.02 (0)	26.72	10 $\rightarrow$20	0 $\rightarrow$9, 100	-
5-mix diag-cov	24.16 (5)	27.08 (4)	24.83	5 $\rightarrow$13	0 $\rightarrow$4, $\geq$ 14	-
8-mix diag-cov	24.63 (4)	27.28 (3)	24.85	4 $\rightarrow$7.25	0 $\rightarrow$2, $\geq$ 8	3
16-mix diag-cov	24.52 (2.5)	27.89 (2.1)	25.19	2.2 $\rightarrow$5.5	0 $\rightarrow$1.9, $\geq$ 5.6	2.0-2.1
32-mix diag-cov	24.23 (1.4)	27.02 (0)	25.19	1.2 $\rightarrow$2.4	0 $\rightarrow$1.1, $\geq$ 2.5	-
64-mix diag-cov	23.90 (0.7)	27.02 (0)	25.19	0.7 $\rightarrow$1.2	0 $\rightarrow$0.6, $\geq$ 1.3	-
128-mix diag-cov	23.76 (0.5)	27.02 (0)	25.19	0.5 $\rightarrow$0.7	0 $\rightarrow$0.4, $\geq$ 0.8	-

Table 25: *Example 2: Summary of the results when using the standard BIC formula to choose the 'best' alternative from the CVOS. The inputs give DERs of 25.12 and 27.09% respectively.*

Base Model	ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
1 full-Gaussian (s)	32.85	20.48	21.15	32.06	**37.18**	**10.74**	*25.02*
1 diag-covariance (s+d)	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
5-mix GMM diag cov (s+d)	32.85	21.17	21.15	32.06	**37.18**	**10.74**	25.16
1 diag-covariance (s)	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
5-mix GMM diag cov (s)	31.89	20.80	21.15	*31.05*	**37.18**	**10.74**	*24.80*
10-mix GMM diag cov (s)	30.30	19.94	***18.11***	*31.05*	**37.18**	**10.74**	*23.79*
15-mix GMM diag cov (s)	30.30	*19.27*	***18.11***	***29.84***	**37.18**	**10.74**	*23.48*
20-mix GMM diag cov (s)	30.30	*19.27*	***18.11***	*30.18*	**37.18**	**10.74**	*23.53*
25-mix GMM diag cov (s)	30.30	*19.27*	***18.11***	*30.50*	**37.18**	**10.74**	*23.57*
32-mix GMM diag cov (s)	32.84	*19.27*	*20.10*	*30.39*	37.92	**10.74**	*24.42*
Input 1	32.03	20.78	21.40	32.06	37.92	**10.74**	25.12
Input 2	29.26	19.82	20.48	31.56	**37.18**	29.34	27.09
Best Possible from CVOS	26.71	18.43	18.11	29.84	37.18	10.74	22.79
Worst Possible from CVOS	34.58	22.48	23.56	33.78	37.92	29.34	29.44

Table 26: *Example 2: DERs when using the Equal-parameter BIC method to choose the 'best' alternative from the cluster voting output set. Numbers in italics are when the final output is better than either input. Numbers in bold match the best possible number from the CVOS. (s) means static coefficients only, whereas (s+d) means statics and deltas.*

System	ABC	VOA	PRI	NBC	CNN	MNB	TOTAL
Input 1	32.03	20.78	21.40	32.06	37.92	**10.74**	25.12
Input 2	29.26	19.82	20.48	31.56	**37.18**	29.34	27.09
Best Possible from CVOS	26.71	18.43	18.11	29.84	37.18	10.74	22.79
Worst Possible from CVOS	34.58	22.48	23.56	33.78	37.92	29.34	29.44
Random	*28.30*	*19.76*	23.18	***29.84***	**37.18**	**10.74**	*24.30*
All the same	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
All different	*27.66*	19.82	20.74	31.56	37.92	29.34	27.02
†Standard BIC (full-cov)	30.30	19.94	*19.15*	32.06	37.92	**10.74**	*24.26*
†Standard BIC (diag-cov)	30.30	19.94	***18.11***	*31.05*	37.92	**10.74**	*23.90*
†Standard BIC (128mix GMM)	*27.66*	20.78	*19.15*	*30.83*	**37.18**	**10.74**	*23.76*
Equal-param BIC (full-cov)	32.85	20.48	21.15	32.06	**37.18**	**10.74**	*25.02*
Equal-param BIC (diag-cov)	33.63	20.78	21.15	32.06	**37.18**	**10.74**	25.19
Equal-param BIC (15mix GMM)	30.30	*19.27*	***18.11***	***29.84***	**37.18**	**10.74**	*23.48*

Table 27: *Example 2 : Summary of Key DERs. Numbers in italics are when the final output is better than either input. Numbers in bold match the best possible number from the CVOS. † Results are for the optimal α value in the standard BIC formulation.*

References

Ajmera, J. and Wooters, C. (2003). A Robust Speaker Clustering Algorithm, *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, St Thomas, US Virgin Islands, pp. 411–416. November 2003

Ajmera, J., Bourlard, H. and Lapidot, I. (2002). Improved Unknown-Multiple Speaker Clustering Using HMM, *Technical Report RR-02-23*, IDIAP Research, `http://www.idiap.ch/publications`. September 2002.

Ajmera, J., McCowan, I. and Bourlard, H. (2004). Robust Speaker Change Detection, *IEEE Signal Processing Letters* To appear.

Bath University WWW site (2004). Bell Numbers, `http://students.bath.ac.uk/ns1tcll/bell.html`.

Bell, E. T. (1934). Exponential Numbers, *American Mathematical Monthly* **41**: 411–419.

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

Campbell, J. P., Reynolds, D. A. and Dunn, R. B. (2003). Fusing High- and Low-Level Features for Speaker Recognition, *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, pp. 2665–2668. September 2003.

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

Chen, S. S., Eide, E., Gales, M. J. F., Gopinath, R. A., Kanvesky, D. and Olsen, P. (2002). Automatic Transcription of Broadcast News, *Speech Communication* **37**: 69–87.

Evermann, G. and Woodland, P. C. (2000). Posterior Probability Decoding, Confidence Estimation and System Combination, *Proc. 2000 Speech Transcription Workshop*, College Park, MD. May 2000.

Fiscus, J. G. (1997). A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Santa Barbara, CA. December 1997.

Gauvain, J. L. and Barras, C. (2003). Speaker Diarization, *Proc. Spring 2003 Rich Transcription Workshop (RT-03s)*, Boston, MA. May 2003.

Gauvain, J. L., Lamel, L., Schwenk, H., Adda, G., Barras, C., Chen, L., Lefevre, F., Meignier, S. and Messauodi, A. (2004). Summary of Progress at LIMSI, *EARS Mid-Year Meeting*, Vienna, VA. February 2004.

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

Johnson, S. E., Jourlin, P., Spärck Jones, K. and Woodland, P. C. (2001). Spoken Document Retrieval for TREC-9 at Cambridge University, *in* E. M. Voorhees and D. K. Harman (eds), *The Ninth Text REtrieval Conference (TREC-9)*, number SP 500-249, Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, pp. 117–126.

Kinnunen, T., Hautamäki, V. and Fränti, P. (2003). On the Fusion of Dissimilarity-Based Classifiers for Speaker Identification, *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, pp. 2641–2644. September 2003.

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

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

Moraru, D., Meignier, S., Besacier, L., Bonastre, J.-F. and Magrin-Chagnolleau, I. (2003). The ELISA Consortium Approaches in Speaker Segmentation during the NIST 2002 Speaker Recognition Evaluation, *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 2, Hong Kong, pp. 89–92. April 2003.

Moraru, D., Meignier, S., Besacier, L., Bonastre, J.-F. and Magrin-Chagnolleau, I. (2004). The ELISA Consortium Approaches in Broadcast News Speaker Segmentation during the NIST 2003 Rich Transcription Evaluation, *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, To appear. May 2004.

Nguyen, P. and Junqua, J. C. (2003). PSTL's Speaker Diarization, *Proc. Spring 2003 Rich Transcription Workshop (RT-03s)*, Boston, MA. May 2003.

NIST (2000+). Benchmark Tests : Speaker Recognition Evaluations, `http://www.nist.gov/speech/tests/spk/`.

NIST (2002). The NIST Year 2002 Speaker Recognition Evaluation Plan, `http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrec-evalplan-v60.pdf`. 27th February 2002.

NIST (2003a). 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.

NIST (2003b). 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.

Reynolds, D. A. and Torres-Carrasquillo, P. (2004). MIT-LL Diarization: Progress, Plans and Issues, *EARS Mid-Year Meeting*, Vienna, VA. February 2004.

Reynolds, D. A., Andrews, W., Campbell, J. P., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D. and Xiang, B. (2003). The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition, *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 4, Hong Kong, pp. 784–787. April 2003.

Sloane, N. J. A. (2003). (editor) The On-Line Encyclopedia of Integer Sequences, `http://www.research.att.com/~njas/sequences/`.

Tranter, S. E. and Reynolds, D. A. (2004). Speaker Diarisation for Broadcast News, *Proc. Odyssey Speaker and Language Recognition Workshop*, Toledo, Spain, To appear. June 2004.

Tranter, S. E., Yu, K., Reynolds, D. A., Evermann, G., Kim, D. Y. and Woodland, P. C. (2003). An Investigation into the Interactions between Speaker Diarisation Systems and Automatic Speech Transcription, *Technical Report CUED/F-INFENG/TR-464*, Cambridge University Engineering Department. October 2003.

Weisstein, E. W. (2004). Stirling Transform, `http://mathworld.wolfram.com/StirlingTransform.html`.

Woodland, P. C., Hain, T., Johnson, S. E., Niesler, T. R., Tuerk, A., Whittaker, E. W. D. and Young, S. J. (1998). The 1997 HTK Broadcast News Transcription System, *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, pp. 41–48. February 1998.

Note many of the Cambridge University publications are available from
`http://mi.eng.cam.ac.uk/reports`
and see also the CUED EARS-project reference page
`http://mi.eng.cam.ac.uk/research/projects/EARS/references.html`

Publications from the Rich Transcription Workshops can be found through
`http://www.nist.gov/speech/publications`