TWO-WAY CLUSTER VOTING TO IMPROVE SPEAKER DIARISATION PERFORMANCE

S. E. Tranter

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

ABSTRACT

A cluster-voting scheme is described which takes the output from two speaker diarisation systems and produces a new output which aims to have a lower speaker diarisation error rate (DER) than either input. The scheme works in two stages, firstly producing a set of possible outputs which minimise a distance metric based on the DER and secondly voting between these alternatives to give the final output. Decisions where the inputs agree are always passed to the output and those where the inputs differ are re-evaluated in the final voting stage. Results are presented on the 6-show RT-03 Broadcast News evaluation data, showing the DER can be reduced by 1.64% and 2.56% absolute using this method when combining the best two Cambridge University and the best two MIT Lincoln Laboratory diarisation systems respectively.

1. INTRODUCTION

Speaker diarisation is the task of automatically segmenting audio data and providing speaker labels for the resulting regions of audio. This has many applications such as enabling speakers to be tracked through debates, allowing speaker-based indexing of databases, aiding speaker adaptation in speech recognition and improving readability of transcripts. The speaker labels produced are 'relative' (such as 'spkr1') in that they show which segments of audio were spoken by the same speaker, and *do not* attempt to give a true identity (such as 'David Koppel') of the speaker.

The Rich Transcription diarisation evaluations[1, 2] provide a framework to analyse the performance of such systems on Broadcast News (BN) data. A Diarisation Error Rate (DER) is defined which considers the sum of the missed, false alarm and speaker-error rates after an optimal one-to-one mapping of reference and hypothesis speakers has been performed. (This mapping is necessary to associate the 'relative' speaker labels from the hypothesis to the 'true' speaker labels in the reference, and is chosen to maximise the sum over all reference speakers of the time that is jointly attributed to both the reference and the corresponding mapped hypothesis speaker. See [2] for more details).

In general the error tends to be dominated by the speaker-error components, and in particular, since the DER is time-weighted, decisions about the predominant speakers can have a very large effect. Therefore it is desirable to be able to 'check' such decisions, perhaps by combining information from different sources.

Several methods of combining aspects of different diarisation systems have been tried, for example the 'hybridization' or 'piped' CLIPS/LIA systems of [3, 4] and the 'Plug and Play' CUED/MIT-LL system of [5] which both combine components of different systems together. A more integrated merging method is described in [4], whilst [3] describes a way of using the 2002 NIST speaker segmentation error metric to find regions in two inputs which agree and then uses these to train potentially more accurate speaker models. These systems are interesting but tend to place some restriction on the systems being combined which we would like to remove.

This paper describes a cluster-voting scheme which takes the output from any two different diarisation systems as input and tries to produce a new output which is better than either input. This is achieved by reproducing the decisions on which the inputs agree whilst allowing an external judge to decide in the case of conflict. The paper is arranged as follows, section 2 describes the theory behind the scheme, section 3 describes the data used in the experiments, section 4 gives the experimental results and conclusions are offered in section 5.

2. THE CLUSTER VOTING SCHEME

The cluster voting scheme is illustrated in Figure 1. The process is divided into two steps. Firstly the inputs are compared and the non-conflicting parts are passed straight to the output. A Cluster Voting Metric (CVM) based on the DER is then used on the remaining data to find sets of 'optimal' speaker labellings under the CVM and these alternatives are output in the Cluster Voting Output Set (CVOS). The second stage consists of choosing the final output from the CVOS using an external judge. These stages are discussed further in sections 2.1 and 2.2 respectively.

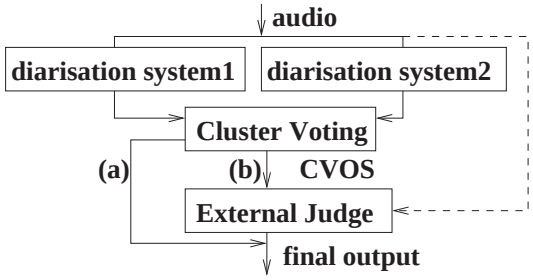

Fig. 1. Cluster Voting Architecture: The cluster voting scheme sends decisions on which the inputs agree straight to the output (a) whilst the disagreements are placed in the CVOS and resolved by an external judge (b).

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.

2.1. Generating the Cluster Voting Output Set

The stages of the CVOS generation are as follows:

1. Form a common base segmentation from the two inputs, such that no input speaker changes in any base segment.
2. Resegment the data to 'tie' together those base segments which have the same speaker label in input-1 *and* the same speaker label in input-2.
3. Pass non-conflicting resegments directly to the final output.
4. Generate independent supergroups of the remaining resegments.
5. Form sets of speaker labellings which minimise the cluster voting metric (CVM) for each supergroup.

These steps are explained in more detail below and illustrated in Figure 2.

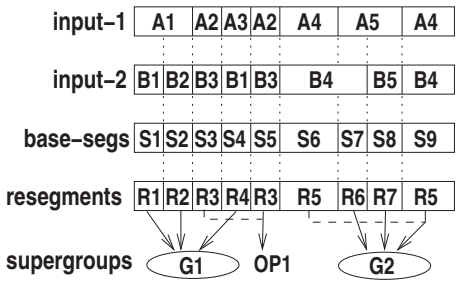

Fig. 2. First steps in the CVOS generation. Base segments are formed which contain no speaker change in either input. Resegments are formed which tie together those segments with a single speaker label in input-1 and a single speaker label in input-2. Here R3={S3/S5} and R5={S6/S9}. Non-conflicting resegments are passed to the output whilst the others are grouped into independent supergroups. Here {A2/B3} do not overlap with any other speakers and hence form output OP1, whilst {R1/R2/R4}≡{A1/A3/B1/B2} form G1 and {R5/R6/R7}≡{A4/A5/B4/B5} form G2. This reduces the number of *all possible* clusterings of the base-segments for this example from $\text{Bell}(9) = 21,147$ to $2 \times \text{Bell}(3) = 10$.

Firstly a common base segmentation is made by dividing up the data as necessary to ensure there is no speaker change for either input in any base segment.[1] This simplifies the calculation of overlap time between the different clusterings, since each segment can be represented by its duration and input speaker-ids alone.[2]

Resegments are then formed by associating groups of base segments together which have a common speaker label in input-1 and a common speaker label in input-2. This is effectively the same as making new speaker labels which are the concatenation of the two input labels for each segment, and accumulating the durations for each of these *new* resegment speaker labels. No other restriction (such as one of temporal adjacency) is used to define the resegments, and no decision is made about the final speaker label of their base-segments other than to ensure they will have the same speaker-id in the final output. This usually significantly reduces the number of 'segments' to be considered - thus reducing the computational complexity. For example, this reduces the number of 'segments' from 869 to 162 in Expt. 1 (section 4.1) and from 557 to 303 in Expt. 2 (section 4.2).

Any resegments which are non-conflicting (in that there is no base segment which has the same input speaker-id as the resegment for either input, but which is not itself in the resegment) are passed directly to the output with a single unique speaker-id for the resegment. This reduces the complexity further (removing 71 out of 162 resegments in Expt. 1 and 48 out of 303 in Expt. 2).

The speaker-ids of some groups of resegments may be independent of other groups. For example, if the clustering has been done gender dependently, there will never be any overlap between male and female speaker-ids and thus these two groups can be treated separately. It is possible to find these independent *'supergroups'* of resegments automatically (see [6] for details) and subsequently processing them separately reduces the computational complexity further. For the example in Figure 2 this reduces the number of all possible clusterings of resegments from Bell(6)=203 to 2xBell(3)=10.

Finally, sets of speaker labellings are formed which minimise the Cluster Voting Metric (CVM) for each supergroup. This metric is the sum of the DERs from the output to both inputs and reduces to the same as maximising the sum of the overlap between the output speaker labels and the speaker labels of both inputs under the optimum one-to-one speaker mappings performed in the DER calculation, if there is a common input segmentation. When the difference between the inputs is relatively small, it is possible to generate all possible speaker labels for the resegments and score them to find the CVOS exhaustively (see [6] for details), but it is also possible to generate the CVOS members directly from the optimum speaker mapping between the two inputs.

Mapping the inputs leads to a speaker mapping which is either *'good'* (the input-1 speaker is mapped to the input-2 speaker), *'bad'* (the input-1 speaker is mapped to a different input-2 speaker) or *'null'* (either the input-1 or the input-2 speaker is not mapped). A single unique speaker-id can be assigned to each 'good' mapping. A 'bad' mapping produces two alternatives, namely those corresponding to the mapped input-1 id or the input-2 id. The 'null' mappings are more complicated but usually give rise to two alternatives. The first is from the mapped input-id which is not null, whilst the second is either a new (unseen) id or is from the input-id which is mapped to null, depending on whether the id has already been seen in that particular member of the CVOS. Further alternatives also arise if a supergroup contains both input-1 and input-2 speaker-ids which are not mapped.

The overall CVOS is formed from the outputs from all the supergroups. Further simplifications can be applied when generating the final CVOS. For example, giving a single speaker-id to the resegments in supergroups of less than a critical duration can reduce the complexity without detrimentally affecting the output, since the DER is time-weighted and thus supergroups with a small duration seldom impact on the final score. (See [6] for an example). Alternatively, if the two inputs are very different, one can allow only the labellings corresponding to the two inputs to be passed into the CVOS for a given supergroup thus effectively enforcing an upper-bound on the size of CVOS for each supergroup, hence preventing complexity problems in the subsequent judging stage.

[1] Diarisation systems are assumed not to output overlapping speakers.

[2] For both experiments reported in this paper the two inputs were generated using different clustering schemes after the Cambridge or MIT-LL base segmentation respectively, so this stage was not necessary.

2.2. Judging the Cluster Voting Output Set

Once the CVOS has been generated a method of choosing the final output must be defined. For the two input case, using confidence scores on the inputs would simply result in the input with the highest confidence being output directly (although this method could be used when there are more than two inputs). Alternative strategies include simple schemes such as assigning the resegments in a supergroup all the same or all different speaker-ids depending on the size of the supergroup, or just picking a member of the CVOS at random, or more mathematically based methods such as using the Bayes Information Criterion (BIC). Here two BIC-based schemes are investigated.

2.2.1. Standard BIC Model Selection

The Bayes Information Criterion gives a log likelihood of the data, $\mathcal{L}$, which is penalised in proportion to the number of parameters in the model:[7]

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

where $\#M$ is the number of free parameters, N the number of data points and α the tuning parameter. If K clusters are each modelled using Gaussian(s) of dimension d which have N_i frames and a covariance S_i then maximising (1) is the same as minimising:

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

where $P = [0.5d(d+1) + d]$ for a full covariance Gaussian, or $[2dG + G - 1]$ for a G-mixture diagonal covariance GMM.

For a given member of a supergroup CVOS, a model is made for each output speaker-id. The BMIN value for this set of models given the data segments is then calculated. The final output chosen for the supergroup is the CVOS member which produces the lowest BMIN. The overall output is then simply the concatenation of the final outputs from the (independent) supergroups.

2.2.2. Equal Parameter BIC Model Selection

An alternative BIC-based strategy removes the need for the tunable α parameter by ensuring all the model sets being judged contain the same number of free parameters[8]. For example, consider whether to split a parent cluster into two children. A model is built for the children with M_{c1} and M_{c2} free parameters respectively. The parent model is then built using $M_p = M_{c1} + M_{c2}$ parameters. The choice between children and parent thus reduces to taking the one that gives the highest likelihood for the data. Here a model is built for each speaker-id as before, but the number of parameters is made proportional to the number of constituent resegments.

3. DATA USED IN EXPERIMENTS

The experiments reported in this paper were conducted on the 6-show 2003 evaluation data *(bneval03)* used in the English Broadcast News RT-03 Rich Transcription evaluations.[1] It consists of 30 minute extracts from 6 different US news shows broadcast in February 2001. Two of these are from radio sources, namely VOA and PRI, whilst four are TV sources, namely NBC, ABC, MNB and CNN. (see [9] for more details) The diarisation references were generated using the rules described in [10] using forced alignments provided by the LDC and with 0.3s of silence smoothing applied, and no collars were used during scoring.

4. EXPERIMENTAL RESULTS

4.1. Experiment 1: Using CUED's Diarisation Systems

The two best diarisation systems from Cambridge University in December 2003 were used as inputs to the cluster voting. They both use the CUED RT-03s BN segmentation which is based on a GMM speech/music classifier, phone recogniser and smooth/clusterer, followed by a top-down clustering stage using PLP coefficients and the arithmetic harmonic sphericity distance measure. Input-1 uses a BIC-based stopping criterion, whilst Input-2 uses one based on node-cost. Further details can be found in [5].

The DERs on the `bneval03` data were 25.12% and 27.09% respectively although Input-2 was the better system for 5 out of the 6 shows. The DER is 24.16%(28.05%) if the best(worst) input is taken independently for each show. Since the systems were relatively similar, it was possible to exhaustively search all the combinations of the CVOS to find the best and worst possible choice for each show[3] which leads to a DER of 22.79% and 29.44% respectively. These results are given in Table 1 along with a summary of those from using the BIC-based judging schemes. More comprehensive results can be found in [6].

The results show that the standard BIC technique can be used to reduce the DER to 23.76%, a 1.36% absolute improvement over the best input, whilst the equal-parameter BIC technique can reduce the DER to 23.48%, a 1.64% absolute reduction. Further experiments reported in [6] show there is a reasonable range in both α value (where applicable) and parameterisation that give an improvement over both inputs.

System	ABC	VOA	PRI	NBC	CNN	MNB	ALL
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 Input	29.26	19.82	20.48	31.56	37.18	10.74	(*24.16*)
Worst Input	32.03	20.78	21.40	32.06	37.92	29.34	(28.05)
Best CVOS	*26.71*	*18.43*	*18.11*	*29.84*	37.18	10.74	(*22.79*)
Worst CVOS	34.58	22.48	23.56	33.78	37.92	29.34	(29.44)

Final Output after Judging								
BIC	Cov	ABC	VOA	PRI	NBC	CNN	MNB	ALL
†St	full	30.30	19.94	*19.15*	32.06	37.92	10.74	*24.26*
†St	diag	30.30	19.94	*18.11*	*31.05*	37.92	10.74	*23.90*
†St	16mix	31.72	20.78	*19.15*	32.06	37.18	10.74	*24.52*
†St	128mix	*27.66*	20.78	*19.15*	*30.83*	37.18	10.74	*23.76*
EP	full	32.85	20.48	21.15	32.06	37.18	10.74	*25.02*
EP	diag	33.63	20.78	21.15	32.06	37.18	10.74	25.19
EP	15mix	30.30	*19.27*	*18.11*	*29.84*	37.18	10.74	*23.48*

Table 1. Example 1 : DERs when using cluster voting on the CUED diarisation systems. Numbers are presented for the standard (St) and Equal-parameter (EP) BIC judging schemes using different covariance representations. Numbers in italics are when the final output is better than either input. Numbers in brackets are from combining the separate outputs from each show. †Results reported for the Standard BIC technique use the optimal α value.

[3] Although the speaker-ids of the supergroups are independent for both the inputs and the output of the cluster voting scheme, there is no guarantee that this is also the case for the true reference speakers, so when scoring against the reference, the supergroups may no longer be treated independently thus dramatically increasing the complexity - from 266 possibilities to 24,992 in Experiment 1.

4.2. Experiment 2: Using MIT-LL's Diarisation Systems

The experiment was repeated using the two-best MIT-LL systems of February 2004.[4] Input-1 is identical to the system described in [5] except that a single full-covariance Gaussian is used in the agglomerative clustering stage. Input-2 used the same segmentation and speech/non-speech detection stages, but the clustering used a system where speakers are represented by their distance to a set of proxy models.[11] These models were created by adapting a GMM trained on the entire audio file to each speech segment in turn. The segments are then represented by a vector of normalised scores against the proxy models, and the final clustering uses a Euclidean distance and BIC-style stopping criterion.

The inputs scored 21.38% and 20.59% respectively, but the standards were considerably different for the individual shows. The scores from taking the best (worst) input per show separately gave 18.38% (23.59%). The inputs were very different in places, one supergroup having 39 resegments and thus potentially $> 10^9$ members of its CVOS, so supergroups of more than 12 resegments (generally ˜500 CVOS members) simply passed the two input possibilities directly to the final CVOS to reduce the complexity. Similarly, for complexity reasons, the best (worst) possible scores in the CVOS have been replaced by the best (worst) show scores seen in all inhouse experiments on the CVOS, which gave a DER of 16.0% (25.56%).

The results, given in Table 2, again show that all the experiments give a lower DER than either input except for the equal-parameter (EP) BIC technique using a single diagonal covariance per resegment. The best DER for the EP BIC scheme was 20.33%, a 0.26% absolute improvement over the best input; whilst the best DER using the standard BIC scheme was 18.03%, a 2.56% absolute reduction over the best input.

System	ABC	VOA	PRI	NBC	CNN	MNB	ALL
Input 1	30.18	22.03	11.99	26.23	32.91	8.99	21.38
Input 2	28.41	16.96	15.35	24.92	22.77	18.83	20.59
Best Input	28.41	16.96	11.99	24.92	22.77	8.99	(*18.38*)
Worst Input	30.18	22.03	15.35	26.23	32.91	18.83	(23.59)
Best Seen	*23.60*	16.96	*8.96*	*17.93*	22.77	*8.54*	(*16.00*)
Worst Seen	35.72	22.27	15.35	33.22	33.62	18.83	(25.56)

Final Output after Judging								
BIC	Cov	ABC	VOA	PRI	NBC	CNN	MNB	ALL
†St	full	*24.31*	19.97	12.75	*19.34*	22.77	17.50	*19.09*
†St	diag	*24.86*	19.86	*10.81*	*19.34*	22.77	12.84	*18.03*
†St	16mix	*23.60*	19.86	*11.78*	*20.37*	22.77	12.28	*18.10*
†St	128mix	*26.34*	18.21	11.99	*24.26*	32.91	8.99	*19.75*
EP	full	31.34	17.79	*11.54*	25.16	33.62	8.99	*20.53*
EP	diag	31.34	18.21	12.06	30.78	33.62	8.99	21.54
EP	16mix	*27.01*	19.66	*10.45*	26.88	33.62	8.99	*20.33*

Table 2. Example 2 : DERs when using cluster voting on the MIT-LL diarisation systems. Numbers are presented for the standard (St) and Equal-parameter (EP) BIC judging schemes using different covariance representations. Numbers in italics are when the final output is better than either input. Numbers in brackets are from combining the separate outputs from each show. †Results reported for the Standard BIC technique use the optimal α value.

[4]Thanks to Doug Reynolds for providing the MIT-LL system outputs.

5. CONCLUSIONS

This paper has presented a cluster-voting scheme designed to reduce the diarisation error rate (DER) by combining information from two different diarisation systems. Results on the RT-03 BN evaluation data show the DER can be reduced by 1.64% and 2.56% absolute over the best input when combining the best two systems from Cambridge University and the best two systems from MIT Lincoln Laboratory respectively.

6. REFERENCES

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

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

[3] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, "The ELISA Consortium Approaches in Speaker Segmentation during the NIST 2002 Speaker Recognition Evaluation," in *Proc. ICASSP*, Hong Kong, April 2003, vol. 2, pp. 89–92.

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

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

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

[7] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition," in *Proc. ICASSP*, Seattle, WA, May 1998, vol. 2, pp. 645–648.

[8] J. Ajmera, H. Bourlard, and I. Lapidot, "Improved Unknown-Multiple Speaker Clustering Using HMM," Tech. Rep. RR-02-23, IDIAP Research, Sept. 2002.

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

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

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