The Cambridge University March 2005 Speaker Diarisation System

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

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

Abstract

This paper describes the speaker diarisation system developed at Cambridge University in March 2005. This system combines techniques used successfully in our previous speaker diarisation systems with an additional second clustering stage based on state-of-the-art speaker identification methods. Several strategies for using the new system are investigated and the final system gives a diarisation error rate of 6.9% on the RT-04 Fall diarisation evaluation data when processing all the test data together or 8.6% when processing the test data shows independently.

1. Introduction

Audio diarisation is the task of automatically segmenting an input audio stream into acoustically homogeneous segments and attributing them to sources. In general, these sources can include particular speakers, music, background noise sources and other source/channel characteristics. In the NIST Rich Transcription evaluations[1] within the DARPA EARS program, the task is limited to speaker diarisation; namely providing a list of 'who spoke when' throughout some audio data.

Speaker diarisation 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 automatic transcripts.

In this paper we describe the speaker diarisation system developed at Cambridge University in March 2005. This draws on techniques used in our previous diarisation system[2] but includes several modifications to the core components and incorporates a new additional clustering stage. The final system gives a diarisation error rate (DER) of 6.9% on the RT-04 Fall diarisation evaluation data when processing all the test data together, or 8.6% when processing each test data show independently. This compares well with other state-of-the-art diarisation systems [3].[1]

The paper is structured as follows. Section 2 describes our baseline diarisation system. The techniques used to improve the system are described in section 3. Section 4 describes the experimental set up and results. Finally conclusions are offered in section 5.

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.

[1]Note that the RT evaluation rules specify that the test shows must be processed in chronological order.

2. Baseline Diarisation System

The baseline system consists of the three main components typically found in most canonical speaker diarisation systems, namely speech detection, speaker change point detection and speaker clustering [4]. The system is illustrated in Figure 1.

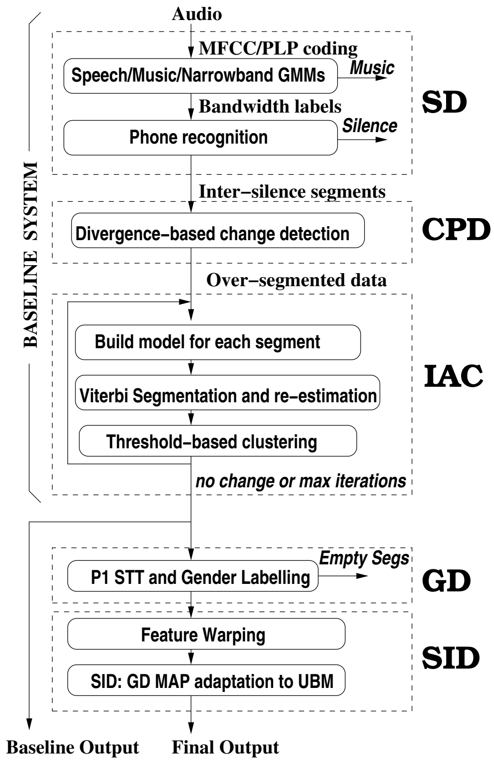

Figure 1: The diarisation system architecture. The baseline system uses the Speech Detection (SD), Change Point Detection (CPD) and Iterative Agglomerative Clustering (IAC) stages, whilst the March 2005 system also includes a Gender Determination (GD) and SID-like additional clustering stage (SID).

In the speech detection (SD) stage, the speech signal is partitioned into regions of wideband speech (S), speech with music (MS), narrowband speech (T) and music only (M) using a GMM classifier incorporating an MLLR adaptation stage. The MS regions are relabelled as S whilst the M portions are discarded. Wideband and narrowband data are subsequently treated independently. A phone recogniser which has 45 context independent phone models per gender plus a silence model with null language model is then run for each bandwidth. Silence portions larger than 1 second are discarded and portions of speech between these silences form the new segments.

The change point detector (CPD) finds potential changes in audio characteristics within each segment using the symmetric divergence (KL2) distance metric between two adjacent sliding windows of 1 second length. A single diagonal covariance Gaussian is used for each window and the distance threshold is chosen to over-segment the data.

The next stage uses an iterative agglomerative clustering (IAC) scheme similar to [5]. A single Gaussian model is built for each segment and the likelihood change for each potential merge of segments is calculated. The merge with the smallest likelihood loss is performed and the statistics are recalculated. This is repeated until the potential likelihood loss on merging reaches a certain threshold. These new models are then used to resegment the data using a Viterbi decode. This whole process is repeated until the segmentation converges or a maximum number of iterations are reached.

The baseline system used up to 6 iterations with both diagonal covariance and then full covariance models. The introduction of the diagonal covariance stage was necessary due to the large number of small segments coming from the CPD stage, but the thresholds were chosen so most of the clustering occurred in the full covariance stage. The likelihood thresholds were phased in, using a smaller value for the first iteration of each stage to reduce the number of merges carried out in any one iteration. A minimum length of 0.6s was enforced on the segments during the IAC stage.

The baseline system for this paper uses a 'local' BIC criterion (as described in e.g. [6]) for both the stopping decision and the ordering of merges, and updates the statistics assuming that the data in the cluster has been concatenated. Different methods such as centroid or furthest neighbour clustering, and using a constant threshold for ordering and/or stopping were tried but did not perform consistently better.

3. The March 2005 Diarisation System

The LIMSI RT-04F diarisation evaluation system showed significant performance improvements by incorporating an additional clustering stage which employed state-of-the-art speaker identification (SID) techniques [7]. We therefore investigated incorporating these ideas into our system, as illustrated in Figure 1. Changes were also made to the CPD algorithm which in turn meant that the diagonal covariance IAC stage could be removed. These changes are now discussed further.

3.1. Gender Determination

The final clustering stage is done gender-dependently. To provide the gender labels, each segment was run through the first pass of the CUHTK RT-03s ASR system [8]. The empty segments were discarded and a forced alignment with gender-dependent models was used to give the final gender labels as was done in [2].

3.2. Speaker Identification (SID) Clustering

A further agglomerative clustering stage was added. Each gender/bandwidth combination was processed separately. Maximum A Posteriori (MAP) adaptation (mean-only) was applied towards each cluster from the appropriate gender/bandwidth Universal Background Model (UBM). Feature warping as described in [9] using a sliding window of 3 seconds was applied to help reduce the effect of the acoustic environment.

The cross likelihood ratio (CLR) between any two given clusters is defined:[7]

$$CLR(c_i, c_j) = \log\left(\frac{L(x_i|\lambda_j)\ L(x_j|\lambda_i)}{L(x_i|\lambda_{ubm})\ L(x_j|\lambda_{ubm})}\right)$$

where $L(x_i|\lambda_j)$ is the average likelihood per frame of data x_i given the model λ_j. The pair of clusters with the highest CLR is merged and a new model is created using all the data in the new cluster. The process is repeated until the highest CLR is below a predefined threshold, θ_{CLR}.

3.3. Iterative MAP Adaptation

In MAP adaptation of GMMs, the mean parameter is estimated using the formula:[10].

$$\hat{\mu}_i = \frac{\tau m_i + \sum_{t=1}^{T} c_{it} x_t}{\tau + \sum_{t=1}^{T} c_{it}} = \frac{\tau m_i + \gamma_i \bar{\mu}_i}{\tau + \gamma_i} \tag{1}$$

where c_{it} is the a posteriori probability for the Gaussian mixture component i given the observation x_t. τ is the fixed relevance factor which controls the balance between the observed data and prior (UBM) mean m_i.

It has been shown that multiple iterations of MAP adaptation of a UBM model when used in conjunction with feature warping has a beneficial effect for speaker recognition [11]. This technique, known as iterative MAP (IT-MAP), keeps the prior fixed, but updates the Gaussian posteriors after each iteration using the model from the previous iteration.

We also investigated the effect of using the mean from the previous MAP iteration instead of the prior mean in equation 1 for the second and subsequent iterations, so the second iteration equation becomes:

$$\hat{\mu}_i^{(2)} = \frac{\tau\left(\frac{\tau m_i + \gamma_i^{(1)}\bar{\mu}_i^{(1)}}{\tau + \gamma_i^{(1)}}\right) + \gamma_i^{(2)}\bar{\mu}_i^{(2)}}{\tau + \gamma_i^{(2)}}$$

where the superscript indicates the iteration. Although not strictly within the MAP framework, we refer to this approach as variable prior MAP (VP-MAP). For IT-MAP, if the value of τ is set small initially, the Gaussian posteriors will not vary from iteration to iteration. VP-MAP yields a simple procedure for gradually decreasing the effect of the prior with the increasing number of iterations.

3.4. Building the UBMs

The baseline UBMs were diagonal covariance Gaussian Mixture Models (GMM) and were built using the 1996/7 Hub4 training data. Additional sets of UBMs were built adding half of the reference development data (dating from October 2000 to January 2001), or using the test data itself (in an unsupervised fashion). Both MAP-adapting the baseline GMM, and concatenating the additional data with the original data and rebuilding the GMMs were tried.

3.5. Modifications in CPD

The change point detection (CPD) algorithm was altered to find local maxima in the divergence distance metric between the sliding windows. A left to right search of these peaks was then made removing the smaller of any pairs of neighbouring peaks which occurred within a specified minimum duration. This in turn meant the need for the diagonal covariance iterations in the IAC stage was reduced.

4. Experiments

4.1. Data and Scoring Metric

The experiments reported in this paper use a development set of 24 US broadcast news shows, denoted `dev24` and a 12-show subset of this, denoted `dev12` which was the official RT-04 Fall diarisation development data. Each show is approximately 30 minutes long. The number of speakers in a show varies between 3 and 39. The `dev12` data was recorded in February 2001 and Nov/Dec 2003, whilst the other 12 dev shows, (`dev12comp`), come from October 2000 to January 2001. The eval set is the official RT-04 Fall diarisation evaluation data and consists of 12 shows recorded in December 2003. See [2] for further details.

The main metric of performance is the diarisation error rate (DER). It is the sum of the missed (speech in reference but not in hypothesis), false alarm (speech in hypothesis but not in reference) and speaker error (mapped reference and hypothesised speakers differ) rates of a system when compared to a manually defined reference. The latter term is calculated by matching the hypothesised speakers to reference speakers using a one-to-one mapping which maximises the total overlap between the reference and (corresponding) hypothesis speakers. Further details can be found in [12].

In addition to this metric, we also use a measure of segment (and cluster) impurity, which represents the DER that would be obtained if applying 'oracle' clustering (using the reference speaker information) whilst not splitting any hypothesis segment (cluster). This also includes the miss and false alarm rate and represents the best possible achievable DER given the segmentation (and clustering).

4.2. Effect of Feature Type and Feature Warping

The effect of the type of feature (PLP, MFCC), inclusion of c_0, energy, or just the differentials thereof was investigated. The results showed that PLP features with first differentials and no energy gave the best performance, outperforming MFCCs by around 0.8%. The feature warping itself was found to significantly reduce the DER from 17.6% to 10.8%.

4.3. Type of MAP, τ, and Number of Iterations

The effect of using the two schemes for multiple iterations of MAP, as discussed in section 3.3, is shown in Figure 2 for 2 iterations. The VP-MAP approach outperformed the IT-MAP approach and 2 iterations was found to give better results than 1 or 5 iterations, with the optimal τ being 10 for this case.

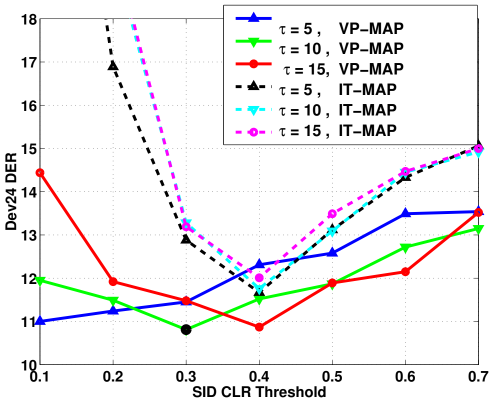

Figure 2: Effect of different MAP approaches and τ values used in deriving the cluster models in the SID stage.

4.4. Processing Narrowband Data

Three methods of processing the automatically labelled narrowband (NB) data in the SID clustering stage, namely (i) passing NB clusters directly to the output, (ii) clustering the NB data using the wideband (WB) coding and (iii) clustering the NB data using the NB coding were tried. There was little difference between the (i) and (iii) schemes, with the latter being slightly better on the `dev24` data, but scheme (ii) gave a degredation of around 1.5% absolute in DER for all θ_{CLR} values. This behaviour is different to what we have previously seen for clustering experiments where using wideband coding is beneficial even for segments labelled as NB, although using a separate θ_{CLR} for the NB case may alleviate this somewhat.

4.5. Building the UBM

Several different methods of building the UBMs were investigated. The results are given in Table 1. The experiments reported in the proceeding sections used a 512 component GMM trained on 6 hours of the training data (K-512). That was replaced by a 1024 component GMM trained on 7.5 hours of source-balanced training data (B-1024). This reduced the DER from 10.7% to 10.4% on the `dev24` data.

The `dev12` data was kept as held-out dev data, and the reference information for the remaining twelve dev shows was used to make new UBMs, to add data from more recently broadcast shows. The `B+D` UBMs simply concatenated the baseline UBM training data, `B`, with the reference dev data, `D`, and then rebuilt the GMMs; whilst the `B→D` models were formed by performing a single iteration of (mean-only) MAP adaptation, with $\tau = 20$ from the `B` models to the `D` data. This reduced the DER on the `dev12` data from 8.6% to 8.4% and on the eval data from 9.9% to 9.5%.

The previous experiment was repeated adding the test data itself, `E`, instead of the dev data. No reference information was used, the automatically generated segment boundaries and gender labels being taken for the model building. This does however violate the (somewhat artificial) constraint in the RT evaluations that shows must be treated chronologically.[2] The DER on the `dev24` data was reduced from 10.4% to around 10% but the eval data saw a much greater drop, from 9.9% to around 8%. This is probably because the eval data set is temporally homogeneous (all 12 shows were broadcast within 18 days), and suggests that collecting contemporaneous data around the test shows could be useful whilst requiring little additional cost.

UBM	Dev24	Dev12	Eval	(OptEval)
K-512	10.7	8.7	9.9	(9.2)
B-512	10.7	9.2	8.6	(8.6)
B-1024	10.4	8.6	9.9	(8.9)
B→D-1024	10.0†	8.5	10.4	(8.0)
B+D-1024	10.3†	**8.4**	**9.5**	(8.3)
B→E-1024	10.2	-	7.8	(7.8)
B+E-1024	**9.9**	-	**8.3**	(8.3)

Table 1: Results of different methods of building the UBM. `B` represents the baseline UBM, `D` exploits the `dev12comp` reference dev data and `E` uses the test data (unsupervised). Concatenation (+) and MAP adapted (→) results are given. (OptEval uses the best θ_{CLR} on the eval data.) † Biased due to the inclusion of `dev12comp` in the D model.

[2]Using just the target test show for E was not effective.

4.6. Modification of the CPD

The new CPD algorithm discussed in section 3.5 which enforced a minimum length constraint on the resulting segments was introduced. The results, illustrated in Figure 3, show that this method successfully reduces the segment impurity. Further improvements can be obtained by increasing the size of the sliding windows from 1 to 2 seconds, with a corresponding increase in minimum segment length of 0.5 to 1 second; and using full covariance models on these larger windows. This reduced the speaker error component of the segment impurity after the CPD stage of the full system on the `dev24` data from 1.7% to 0.2% whilst keeping the number of segments around 11,000.

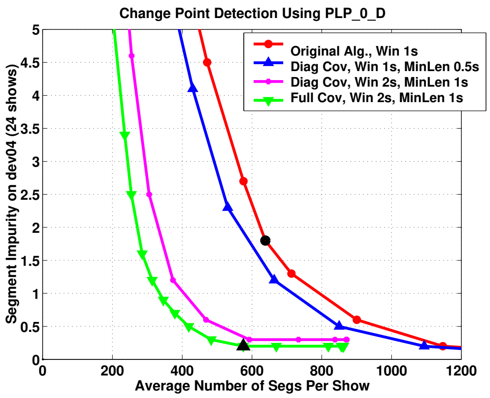

Figure 3: Effect on Segment Impurity (excluding miss and false alarm constituents) with improved change point detection.

4.7. Balancing the Clustering Stages

The new minimum length constraint on the CPD meant that full covariance models could be used throughout the IAC stage. The number of iterations of IAC was reduced to 2, with the tunable parameter, α, set to 0 and then 1.5. Further iterations of the IAC stage were found not to make significant changes in performance. The chosen settings heavily under-clustered the data during the IAC stage (see Table 2), but provided a reasonable starting point for the more successful SID clustering stage.[3] The final results for the complete system are given in Table 3.

IAC	DataSet	MS/FA/SPE/DER	Cluster Imp	Seg Imp
Base	Dev24	1.2/1.1/17.9/20.17	6.59@718	4.36@2363
Base	Eval	0.3/1.1/17.4/18.75	5.04@336	3.63@1072
Final	Dev24	1.2/1.1/41.3/43.55	3.69@1744	3.12@2877
Final	Eval	0.3/1.1/43.3/44.65	2.60@838	2.44@1323

Table 2: Results after the IAC stage. The final system uses a heavily under-clustered IAC stage, giving a much higher DER but lower segment/cluster impurity for the final SID stage.

UBM	Dev24	Dev12	Eval	(OptEval)
B-1024	9.2	7.5	9.3	(8.4)
B→D-1024	9.0†	7.7	9.3	(7.3)
B+D-1024	8.8†	**7.4**	**8.6**	(7.8)
B→E-1024	9.1	-	7.5	(7.5)
B+E-1024	**9.0**	-	**6.9**	(6.8)

Table 3: Results using the new CPD and IAC stages. UBM definitions are the same as Table 1.

[3] Previously α was set to 1 then 3.9.

5. Conclusions

This paper has described the development of Cambridge University March 2005 diarisation system. Motivated by other work, a state-of-the-art SID-like clustering stage was introduced and shown to dramatically cut diarisation error rate. Further modifications and enhancements have been introduced with the final system giving a diarisation error rate of 6.9% on the RT-04 Fall evaluation data when processing all the test data simultaneously, or 8.6% when treating the data sequentially.

6. Acknowledgments

Thanks are due to the LIMSI speaker recognition team, particularly Claude Barras, for providing useful information which helped in the development of this system.

7. References

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

[2] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, and P. C. Woodland, "The Development of the Cambridge University RT-04 Diarisation System," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, November 2004.

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

[4] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," in *Proc. ICASSP*, March 2005, vol. V, pp. 953–956.

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

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

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

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

[9] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2001.

[10] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[11] R. Vogt, J. Pelecanos, and S. Sridharan, "Dependence of GMM Adaptation on Feature Post-Processing for Speaker Recognition," in *Proc. Eurospeech*, September 2003, vol. IV, pp. 3013–3016.

[12] 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.