Cross Likelihood Ratio Based Speaker Clustering Using Eigenvoice Models

D. Wang, R. Vogt, S. Sridharan and D. Dean

Speech and Audio Research Laboratory,
Queensland University of Technology,
Brisbane, Australia

di.wang@student.qut.edu.au, r.vogt@qut.edu.au, s.sridharan@qut.edu.au, ddean@ieee.org

Abstract

This paper proposes the use of eigenvoice modeling techniques with the Cross Likelihood Ratio (CLR) as a criterion for speaker clustering within a speaker diarization system. The CLR has previously been shown to be a robust decision criterion for speaker clustering using Gaussian Mixture Models. Recently, eigenvoice modeling techniques have become increasingly popular, due to its ability to adequately represent a speaker based on sparse training data, as well as an improved capture of differences in speaker characteristics. This paper hence proposes that it would be beneficial to capitalize on the advantages of eigenvoice modeling in a CLR framework. Results obtained on the 2002 Rich Transcription (RT-02) Evaluation dataset show an improved clustering performance, resulting in a 35.1% relative improvement in the overall Diarization Error Rate (DER) compared to the baseline system.

Index Terms: eigenvoice modeling, joint factor analysis, cross likelihood ratio, speaker clustering, speaker diarization

1. Introduction

Speaker clustering, the process of associating segments of speech produced by the same speaker, is commonly performed as one of the final stages within a speaker diarization system. The clustering stage is responsible for associating all speech segments belonging to the same speaker by providing them with the same speaker label. Speaker clustering is commonly regarded as the most crucial step in the final stages of a speaker diarization system.

One of the most popular speaker clustering strategies to date involves the use of a distance metric in conjunction with agglomerative clustering, otherwise known as bottom-up hierarchical clustering. In this approach, clustering is performed by iteratively merging the closest pair of clusters, as determined by some distance metric. The distance metric measures the dissimilarity between the two clusters of interest, and the choice of an appropriate distance metric is essential to the success of the clustering system using this approach. Various distance metrics have been proposed in speaker diarization literature, including the Bayesian Information Criterion (BIC) [1], the symmetric Kullback-Leibler divergence [2] and the Bayes Factor [3], each with varying degrees of success. In the LIMSI broadcast news diarization system [4], the Cross Likelihood Ratio (CLR) is used as a distance metric for agglomerative speaker clustering. The CLR criterion elegantly combines the information present in both clusters of interest with knowledge of the show background model. This system was the top participant in the most recent NIST Rich Transcription broadcast news evaluation, the RT-04F [5], and is used as the baseline system in this paper.

Recently, eigenvoice modeling of speaker segments using Joint Factor Analysis (JFA) techniques have become increasingly popular in speaker recognition literature [6]. Compared to traditional Gaussian Mixture Model (GMM) based approaches, which can potentially suffer from the lack of data caused by short speaker segments resulting in poor quality models, eigenvoice modeling enjoys the advantage of being able to adequately represent a speaker with limited enrollment data. This is achieved by taking advantage of the highly informative prior distribution contained in the speaker models, and using only the most prominent eigenvoices, which account for most of the speaker variability. This greatly reduces the dimensionality and hence the number of parameters that need to be estimated. JFA also has the potential to achieve improved capture of differences in speaker characteristics, through explicit and independent modeling of speaker and channel variations. While earlier work on JFA in the area of speaker recognition have generally focused on the speaker verification task, such as in [7], increasing research efforts are being placed on the application of JFA techniques in the speaker diarization task in recent years. Published speaker diarization systems that use eigenvoice modeling for speaker clustering include the Variational Bayes system, as reported in [8]. Inspired by the pioneering work by Valente [9], which used Variational Bayesian methods for speaker clustering, this system combines the success of factor analysis methods in speaker recognition with the advantages of a Bayesian approach to the diarization problem. A significant reduction in DER was achieved over the baseline agglomerative clustering system using the BIC.

This paper proposes that it would be beneficial to incorporate eigenvoice modeling in a CLR framework for speaker clustering, in order to combine the advantages of the two techniques described above. Section 2 presents an overview of the baseline broadcast news diarization system, which performs speaker clustering using traditional GMM based modeling techniques. Section 3 outlines the theory behind eigenvoice modeling, and shows how eigenvoice modeling techniques can be integrated into the CLR framework. Section 4 presents the result obtained on the RT-02 Evaluation dataset and compares the result to the baseline system, and Section 5 draws some conclusions.

2. Baseline system overview

The baseline system used for comparison in this paper is based on the **c-sid** configuration of the LIMSI broadcast news diarization system [4], which was the top participant in the most recent NIST Rich Transcription broadcast news evaluation, the RT-04F [5]. In the baseline system, the audio is first passed through a speech activity detection stage which separates the audio into speech and non-speech regions. Bayes Factor based

speaker segmentation is then performed to partition the speech regions into homogeneous speaker segments, as described in detail in [10]. This is followed by a Viterbi resegmentation stage which aims to refine the segment boundary locations. The set of speaker segments are then passed to the speaker clustering stages of the system, which aim to merge the segments containing the utterances produced by the same speaker.

Speaker clustering is performed in two separate stages, a Bayes Factor based initial clustering stage, as detailed in [3], followed by a second clustering stage, which uses the CLR criterion with traditional GMM based modeling techniques [4]. Both clustering stages use agglomerative clustering. Due to the lack of data in the initial clustering stage, where speaker segments are relatively short, a multivariate normal distribution is used to model the data, as opposed to a GMM. The initial clustering stage merges only the closest speaker segments and is terminated early, resulting in a set of underclustered nodes, which is passed into the second clustering stage that performs further clustering using more complex models. The performance of the initial clustering stage is hence crucial to the success of the overall diarization system, since correct clustering decisions made in this stage will generate pure, homogeneous clusters with sufficient data to be represented by more complex models in the subsequent clustering stage.

At the end of the initial clustering stage, the segment boundaries are refined once more via Viterbi resegmentation. The refined segments are then passed into the second clustering stage, which completes the clustering process using the CLR as the decision criterion. In this clustering stage, the initial clusters have considerably more data than the individual speaker segments passed into the first clustering stage. GMM's are therefore used to model the more complex distributions of data in each speaker cluster. A show background model, represented by a 128-mixture GMM, is first trained using all speech segments from the whole show. Models for each individual speaker cluster are then obtained via MAP adaptation of the GMM means from the show background model, using data from the relevant cluster of interest. The CLR between each pair of clusters is then calculated and agglomerative clustering is performed, iteratively merging the closest pair of clusters until no more suitable merge candidates can be found. The second clustering stage produces the final diarization output, consisting of a relative, show-internal set of speaker labels and their corresponding start and end times.

3. Incorporating eigenvoice modeling in the Cross Likelihood Ratio framework

This section describes how eigenvoice modeling techniques can be integrated into the CLR framework for speaker clustering. A brief summary of the theory behind eigenvoice modeling is first presented. The CLR criterion as a similarity measure is then introduced, followed by a mathematical derivation showing how eigenvoice modeling can be integrated into the CLR framework. Finally, the implementation details of the new speaker clustering systems are outlined.

3.1. Eigenvoice modeling of speaker segments

As in some traditional speaker clustering approaches, eigenvoice modeling techniques are based around the use of GMM's to model a speaker. Let C be the number of mixture components in the GMM, and F be the dimensionality of the feature vector. From common practice in speaker recognition, only the GMM means are adapted during training. A GMM can therefore be conveniently expressed as a $CF \times 1$ supervector, obtained by concatenating the mean vectors of each mixture component.

In eigenvoice modeling, it is assumed that speaker supervectors have a Gaussian distribution of the form

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}\,, \tag{1}$$

where $\boldsymbol{s}$ represents a randomly chosen speaker segment model, and $\boldsymbol{m}$ is a speaker independent UBM mean supervector obtained by the concatenation of the UBM component mean vectors. $\boldsymbol{V}$ is a $CF \times R$ matrix containing R basis supervectors in the eigenspace, often referred to as eigenvoices. While $R \ll CF$, it is assumed that the most prominent R eigenvoices contained in $\boldsymbol{V}$ is capable of capturing most of the speaker variability. This greatly reduces the dimensionality and hence the number of parameters that need to be estimated, allowing adequate speaker segment models to be constructed from limited enrollment data. $\boldsymbol{y}$ is a hidden $R \times 1$ vector of speaker factors. The speaker variability model is trained such that $\boldsymbol{y}$ follows a standard normal distribution [6].

While the most prominent R eigenvoices have been shown to capture most of the speaker variability, adding a residual term $\boldsymbol{D}\boldsymbol{z}$ to the speaker model has proven beneficial in speaker recognition literature. By providing additional modeling power through the introduction of extra model parameters, the residual term aims to model any residual speaker variations that the speaker factor term fails to take into account. The expression for a given speaker segment model then becomes

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}\,. \tag{2}$$

The results obtained using both (1) and (2) under the CLR framework will be reported in this paper.

3.2. The Cross Likelihood Ratio criterion

The CLR between two clusters, containing data x_i and x_j respectively, is given in [4] as

$$CLR = \frac{1}{n_i}\log\frac{p(x_i|M_j)}{p(x_i|M_B)} + \frac{1}{n_j}\log\frac{p(x_j|M_i)}{p(x_j|M_B)}\,, \tag{3}$$

where n_i and n_j are the number of frames in each cluster, $p(x|M)$ denotes the likelihood of the acoustic frames x given model M, and M_B represents the show background model. This symmetric similarity measure elegantly combines the information present in both clusters of interest with knowledge of the show background model.

In the CLR equation, $\frac{1}{n_i}$ and $\frac{1}{n_j}$ serve as normalization constants, in order to compensate for the different amounts of data present in the clusters of interest. If the speech segments present in the two clusters are produced by the same speaker, $p(x_i|M_j)$ and $p(x_j|M_i)$ should be large, resulting in a large CLR value. Therefore, the larger the CLR, the more evidence that the two clusters should be merged into a single cluster, and vice versa.

3.3. The Cross Likelihood Ratio decision criterion using eigenvoice modeling

In order to describe how eigenvoice modeling can be integrated into the CLR framework, it is useful to first define some notations. Let $\boldsymbol{\Sigma}$ be the covariance of the speaker independent UBM; a $CF \times CF$ diagonal matrix whose diagonal blocks are

Σ_c $(c = 1, \cdots, C)$, where Σ_c is the $F \times F$ diagonal covariance matrix corresponding to the mixture component c. Let $\boldsymbol{N}$, $\boldsymbol{F}$ and $\boldsymbol{S}$ denote the zeroth, first and second order statistics of the speaker segment respectively, as defined in [6]. In the eigenvoice modeling framework, it can be shown [6] that the log likelihood of the acoustic frames x, given model M (in this case, the speaker factors $\boldsymbol{y}$), can be written as

$$\log p(x|M) = \sum_{c=1}^{C} \left(N_c \log \frac{1}{(2\pi)^{\frac{F}{2}} |\Sigma_c|^{\frac{1}{2}}} \right) - \frac{1}{2} \operatorname{tr} (\boldsymbol{\Sigma}^{-1} \boldsymbol{S}) + \boldsymbol{y}^* \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{F} - \frac{1}{2} \boldsymbol{y}^* \boldsymbol{V}^* \boldsymbol{N} \boldsymbol{\Sigma}^{-1} \boldsymbol{V} \boldsymbol{y} \,. \quad (4)$$

This expression can be broken down into two parts. The first two terms are dependent only on the data present in the speaker segment, whereas the last two terms also depend on the speaker model. This expression is not very easy to evaluate in its current form. However, the first two terms conveniently cancel out under the CLR formulation, due to the fact that each ratio making up the CLR rely on the same data. Under the CLR criterion, $\log p(x|M)$ can hence be conveniently implemented as

$$\log p(x|M) = \boldsymbol{y}^* \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{F} - \frac{1}{2} \boldsymbol{y}^* \boldsymbol{V}^* \boldsymbol{N} \boldsymbol{\Sigma}^{-1} \boldsymbol{V} \boldsymbol{y} \,. \quad (5)$$

In order to use this result to construct the CLR as a decision criterion for speaker clustering, one must substitute (5) into each relevant term on the right hand side of (3), using the relevant data and speaker models. The end result is shown in (6) below. The CLR between two clusters i and j can be written as

$$\begin{aligned} CLR = {} & \frac{1}{n_i} \Big[(\boldsymbol{y}_j - \boldsymbol{y}_B)^* \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{F}_i - \frac{1}{2} \boldsymbol{y}_j^* \boldsymbol{V}^* \boldsymbol{N}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{V} \boldsymbol{y}_j \\ & + \frac{1}{2} \boldsymbol{y}_B^* \boldsymbol{V}^* \boldsymbol{N}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{V} \boldsymbol{y}_B \Big] \\ & + \frac{1}{n_j} \Big[(\boldsymbol{y}_i - \boldsymbol{y}_B)^* \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{F}_j - \frac{1}{2} \boldsymbol{y}_i^* \boldsymbol{V}^* \boldsymbol{N}_j \boldsymbol{\Sigma}^{-1} \boldsymbol{V} \boldsymbol{y}_i \\ & + \frac{1}{2} \boldsymbol{y}_B^* \boldsymbol{V}^* \boldsymbol{N}_j \boldsymbol{\Sigma}^{-1} \boldsymbol{V} \boldsymbol{y}_B \Big] \,, \end{aligned} \quad (6)$$

where $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are the enrolled speaker factors for clusters i and j respectively, and $\boldsymbol{y}_B$ is the background speaker factors, enrolled using all speech segments from the whole show. This expression can now be used directly as a decision criterion for speaker clustering.

3.4. System implementation

The theory developed in this paper was tested against the final clustering stage of the baseline system, which uses a traditional GMM based modeling approach with the CLR decision criterion. To ensure a fair comparison, the new systems are identical to the baseline system up until the final clustering stage.

Two separate systems were implemented. An intermediate system was first implemented using eigenvoice modeling techniques to adapt the UBM means for each speaker segment. The adapted supervectors are then converted back to a GMM, and the CLR was evaluated, as in the baseline system. In the final system, eigenvoice modeling of speaker segments was integrated into the CLR framework, using (6) given above. The results obtained by the final system will hence demonstrate the degree of overall success achieved by integrating eigenvoice modeling techniques into the CLR framework; whereas the results obtained by the intermediate system will indicate how much of that success can be attributed to the advantages of eigenvoice modeling over traditional GMM based modeling approaches.

In the new systems, the 512-mixture speaker independent UBM was trained using a total of approximately 5.5 hours of speech data, randomly selected from the 1996 and 1997 HUB4 English Broadcast News Corpus, as well as the 1996 USC Marketplace Corpus. $\boldsymbol{V}$ was trained using utterances from 1165 speakers from the same databases, each of whom have at least 60 seconds of total speech. The large amount of data used to train $\boldsymbol{V}$ ensures a strong, highly informative prior on what the speaker model should look like. To prevent any dominant speakers from being overrepresented, data from all speakers with more than 5000 seconds of total speech were truncated to 5000 seconds for training. 300 principal eigenvoices were used to capture the speaker variability. $\boldsymbol{D}$ was trained using utterances from 30 speakers, also from the same databases, each with approximately 60 seconds of speech. In order to maximise the potential of the residual term to model any speaker variations that the speaker factor term fails to take into account, a disjoint set of speakers was used to train $\boldsymbol{D}$.

In the intermediate system, a show model was first adapted from the speaker independent UBM, using all speech segments from the whole show. Each initial cluster was then enrolled independently from the same UBM, resulting in a mean-adapted supervector for each initial cluster. The cluster models were then converted back to traditional GMM's. Since only the means are adapted, the variances and mixture weights of the mean adapted cluster model is the same as that of the UBM. Agglomerative speaker clustering was then performed using the CLR criterion. As in the baseline system, $p(x|M)$ was calculated using the alignment scores of the acoustic frames with the associated model. In each iteration of the agglomerative clustering process, the CLR was calculated for each pair of potential merge candidates, and the closest pair of clusters merged. This process is repeated until no more suitable merge candidates can be found.

In the final system, eigenvoice modeling techniques were integrated into the CLR framework, using (6) given above. The background model $\boldsymbol{y}_B$ was first enrolled using all speech segments from the whole show. Each initial cluster was then enrolled, the value of (6) calculated between each pair of clusters, and agglomerative clustering performed. Once a merge is performed at the end of each iteration, a new $\boldsymbol{y}$ was enrolled for the combined cluster using the combined data from both merge candidates. This new cluster then becomes a merge candidate in future iterations.

4. Results

This section presents the results of the CLR based clustering approach using eigenvoice models, as obtained on the NIST Rich Transcription 2002 (RT-02) Evaluation dataset, and compares the results to the baseline system. Results obtained with and without the residual term $\boldsymbol{Dz}$ in eigenvoice modeling will be reported for both the intermediate and final systems. The RT-02 Evaluation dataset consists of 6 recorded broadcast news shows, each with a scorable region of approximately 600 seconds.

4.1. Performance evaluation metrics

The results obtained by the new systems will be evaluated using the Diarization Error Rate (DER) measure, as defined in [5]. The DER is the primary performance evaluation metric used in the NIST Rich Transcription Diarization tasks. It can be interpreted as the percentage of the total amount of scorable time that is not attributed to the correct speaker, taking into account

speech detection errors. The DER is calculated via an optimal one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs so as to maximize the total overlap between the reference and mapped hypothesis speakers.

4.2. Diarization results

Table 1 below shows the overall diarization results for each system, without the residual term $\boldsymbol{Dz}$ in the eigenvoice modeling. The "Local" results shown are obtained by using the optimal local stopping threshold for each show in the final clustering stage, whereas the "Global" results are obtained by using the same optimal global threshold across all shows that produces the best average DER, a condition that is consistent with the NIST evaluation protocol [5]. The average result of the 6 shows is calculated based on a time weighted average of the amount of scorable time in each show. As evident from Table 1, there is a 24.1% relative improvement in DER between the baseline system and the intermediate system when evaluated using local thresholds, and a 16.6% improvement when evaluated using global thresholds. This improved performance can be attributed to the use of eigenvoice model adaptation of speaker models compared to traditional GMM based modeling. Comparing the intermediate system to the final system, despite the fact that the local results were slightly worse, a 16.0% relative improvement in DER was achieved using a global stopping threshold, due to the fact that the optimal stopping thresholds were very similar across all shows. This improvement can be attributed to the integration of eigenvoice modeling techniques into the CLR framework.

Table 1: Diarization Error Rates (%) - No Residual Term

Show	Baseline		Intermediate		Final	
	Local	Global	Local	Global	Local	Global
1	12.26	21.31	10.02	13.35	7.54	7.54
2	10.58	10.58	6.51	10.55	6.19	6.51
3	1.19	1.28	0.95	1.28	0.95	0.95
4	14.77	16.05	9.44	11.61	11.61	11.61
5	3.87	6.70	4.42	6.10	4.42	5.30
6	26.18	26.18	20.87	25.59	25.24	25.59
Avg DER	**11.66**	**13.92**	**8.85**	**11.61**	**9.48**	**9.75**

Table 2 shows the overall diarization results for each system, including the residual term $\boldsymbol{Dz}$ in the speaker segment model expression. As expected, the result obtained by both the intermediate and final systems outperformed their counterparts shown in Table 1, due to the additional modeling power introduced by the residual term. The intermediate system achieved relative improvements of 24.9% and 24.6% in DER over the baseline, evaluated using local and global stopping thresholds respectively. The final system achieved a further improvement of 4.0% and 13.8% respectively over the intermediate system. Overall, the final system achieved a 35.1% relative improvement in DER compared to the baseline system, based on the "Global" results.

Table 2: Diarization Error Rates (%) - With Residual Term

Show	Baseline		Intermediate		Final	
	Local	Global	Local	Global	Local	Global
1	12.26	21.31	10.02	11.43	7.54	10.02
2	10.58	10.58	6.51	6.51	6.19	6.51
3	1.19	1.28	0.33	0.65	0.95	0.95
4	14.77	16.05	9.44	11.61	11.61	11.61
5	3.87	6.70	4.42	5.30	4.42	5.30
6	26.18	26.18	20.87	26.34	18.93	18.93
Avg DER	**11.66**	**13.92**	**8.76**	**10.49**	**8.41**	**9.04**

5. Conclusions

This paper proposes the use of eigenvoice modeling techniques with the CLR criterion for speaker clustering within a speaker diarization system. By incorporating eigenvoice modeling into the CLR framework, it was possible to capitalize on the advantages of each technique to produce a robust speaker clustering system which outperforms traditional approaches using GMM based modeling. Results obtained on the RT-02 Evaluation dataset show an improved clustering performance using the proposed approach, leading to a 35.1% relative improvement in the overall diarization performance compared to the baseline system. Through the use of an intermediate system, it was also possible to determine how much of that overall improvement can be attributed to the advantages of using eigenvoice modeling of speaker segments over traditional GMM based modeling approaches, and how much has been contributed by integrating eigenvoice modeling techniques into the CLR framework.

6. Acknowledgements

This research was supported by an Australian Research Council (ARC) Linkage Grant No: LP0991238.

7. References

[1] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," *Broadcast News Transcription and Understanding Workshop*, pp. 127–132, February 1998.

[2] M. Siegler, U. Jian, B. Rag, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," *DARPA Speech Recognition Workshop*, pp. 97–99, 1997.

[3] D. Wang, R. Vogt, and S. Sridharan, "Bayes factor based speaker clustering for speaker diarization," *International Conference on Information Science, Signal Processing and their Applications*, pp. 61–64, 2010.

[4] C. Barras, Z. Xuan, S. Meignier, and J. Gauvian, "Multistage speaker diarization of broadcast news," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1505–1512, 2006.

[5] J. Fiscus, "Fall 2004 Rich Transcription RT-04F evaluation plan," *National Institute of Standards and Technology*, 2004.

[6] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," [Online]. Available: http://www.crim.ca/perso/patrick.kenny/.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[8] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal on Selected Topics In Signal Processing*, vol. 4, no. 6, pp. 1059–1070, December 2010.

[9] F. Valente, "Variational Bayesian methods for audio indexing," Ph.D. dissertation, Eurecom, Sophia-Antipolis, France, 2005.

[10] D. Wang, R. Vogt, and S. Sridharan, "Bayes factor based speaker segmentation for speaker diarization," *Interspeech*, pp. 1405–1408, 2010.