Speaker Attribution of Australian Broadcast News Data

Houman Ghaemmaghami, David Dean, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia
{houman.ghaemmaghami,d.dean,s.sridharan}@qut.edu.au

Abstract

Speaker attribution is the task of annotating a spoken audio archive based on speaker identities. This can be achieved using speaker diarization and speaker linking. In our previous work, we proposed an efficient attribution system, using complete-linkage clustering, for conducting attribution of large sets of two-speaker telephone data. In this paper, we build on our proposed approach to achieve a robust system, applicable to multiple recording domains. To do this, we first extend the diarization module of our system to accommodate multi-speaker (>2) recordings. We achieve this through using a robust cross-likelihood ratio (CLR) threshold stopping criterion for clustering, as opposed to the original stopping criterion of two speakers used for telephone data. We evaluate this baseline diarization module across a dataset of Australian broadcast news recordings, showing a significant lack of diarization accuracy without previous knowledge of the true number of speakers within a recording. We thus propose applying an additional pass of complete-linkage clustering to the diarization module, demonstrating an absolute improvement of 20% in diarization error rate (DER). We then evaluate our proposed multi-domain attribution system across the broadcast news data, demonstrating achievable attribution error rates (AER) as low as 17%.

Index Terms: speaker attribution, diarization, linking, complete linkage, broadcast news.

1. Introduction

The recent developments in speaker modeling and recognition techniques, such as joint factor analysis (JFA) modeling [1] and i-vector speaker modeling [2], have brought about great improvements to the field of speaker diarization [3, 4, 5]. This has motivated the proposal of *speaker attribution* as a recent field of research [4, 5, 6, 7, 8, 9]. Speaker attribution is the process of automatically annotating a typically large archive of spoken recordings based on the unique speaker identities that are present within the analysed archive of recordings, without any prior knowledge of the present speaker identities. This annotation can then be employed to search and index the recording archive based on speaker identity. A typical speaker attribution system can be divided into the two independent modules of speaker diarization and speaker linking [4, 5, 9]. In such a system, the set of recordings are first processed using speaker diarization to ideally extract a set of speaker-homogeneous segments from within each recording [10, 11]. These segments are then passed to the speaker linking module of the attribution system, where they are linked to identify segments belonging to the same speaker identities across multiple recordings [6, 8].

One of the main challenges with speaker attribution is the problem of session variation between the analysed set of recordings. Session variability can degrade the performance of speaker linking when attempting to cluster inter-session segments belonging to the same identity. In our previous work, we demonstrated the erroneous effects of inter-session variability on the tasks of speaker linking and attribution, and proposed the use of JFA modeling to overcome this issue [7]. JFA and i-vector modeling have since been the only speaker modeling techniques employed for conducting attribution [4, 5, 6, 9].

As speaker attribution is often employed to process large sets of data [4, 5, 6], it is of great importance to carry out this process in an efficient manner. The most obvious area for gaining efficiency is the clustering module of attribution. In diarization, clustering is typically based on a computationally expensive, agglomerative merging and retraining scheme [10, 11, 12, 13]. This may not pose a problem to diarization efficiency when processing short recordings, however this is highly inefficient for conducting speaker linking in large datasets. For this reason, van Leeuwen proposed an agglomerative clustering approach, without retraining, for speaker linking [6]. We then proposed a complete-linkage approach to clustering, for both diarization and speaker linking using JFA modeling and cross-likelihood ratio (CLR) scoring, and demonstrated that our complete-linkage clustering approach is more efficient and more accurate than traditional agglomerative clustering with retraining and the method proposed by van Leeuwen [7, 5, 8].

State-of-the-art attribution technology has largely dealt with two-speaker telephone recordings [4, 7, 5, 8], with recent work conducted by Ferras and Bourlard on attribution of meeting room data with poor results [9]. In this paper we extend our previously proposed telephone data attribution system [5], to a robust attribution method applicable to multiple recording domains. To do this, we collected a set of real, and publically available, Australian broadcast news recordings, with the topic of the recordings centered around related events to ensure multiple occurrences of identities across recordings. We then carried out a manual annotation of this dataset to obtain the ground-truth diarization labels for evaluation purposes.

As a common assumption in speaker diarization of telephone recordings [4, 5, 3], our previously proposed diarization module employed a stopping criterion of two speakers for the clustering process. We thus need to modify our diarization module to accommodate recordings with an arbitrary number of unique speaker identities. To do this we propose a CLR threshold stopping criterion for speaker clustering in our baseline diarization module. We justify our choice of this threshold value based on the computation of the CLR metric. We then evaluate this baseline diarization module across the broadcast news data and propose an additional pass of the clustering stage to improve the baseline system. We demonstrate an absolute improvement of 20% in DER over the baseline performance through the application of this additional pass of the clustering stage. We then evaluate our proposed speaker attribution system across the broadcast data to reveal an achievable AER of 17%, given an ideal speaker diarization module.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

2. Speaker modeling and clustering

To carry out robust and efficient speaker attribution of inter-session spoken recordings, we draw from our previous work and employ a JFA speaker modeling approach with session compensation [14, 15]. We compare the modeled speaker segments using the pairwise CLR metric [10]. The pairwise CLR scores are then used to conduct a single stage complete-linkage clustering of the speaker segments without retraining. [5, 8]. This section provides the theory behind JFA speaker modeling, pairwise CLR scoring and complete-linkage clustering.

2.1. JFA speaker modeling

We perform JFA modeling with session compensation using a combined gender universal background model (UBM) [14, 15]. To do this, we introduce a constrained offset of the speaker-dependent, session-independent, Gaussian mixture model (GMM) mean supervector, $\boldsymbol{m}$,

$$\boldsymbol{m}_i(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(s) + \boldsymbol{D}\boldsymbol{z}(s) + \boldsymbol{U}\boldsymbol{x}_i(s), \tag{1}$$

where $\boldsymbol{m}$ is the speaker- and session-independent GMM-UBM mean supervector of dimension $CL \times 1$, with C being the number of mixture components used in the GMM-UBM and L the dimension of the features. $\boldsymbol{x}_i(s)$ is a low-dimensional representation of variability in session i, and $\boldsymbol{U}$ is a low-rank transformation matrix from the session subspace to the GMM-UBM mean supervector space. $\boldsymbol{y}(s)$ is the speaker factors, representing the speaker in a specified subspace with a standard normal distribution [15]. $\boldsymbol{V}$ is a low-rank transformation matrix from the speaker subspace to the GMM-UBM mean supervector space. $\boldsymbol{D}\boldsymbol{z}(s)$ is the residual variability not captured by the speaker subspace, where $\boldsymbol{z}(s)$ is a vector of hidden variables with a standard Gaussian distribution, $N(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$. $\boldsymbol{D}$ is the diagonal relevance maximum *a posteriori* (MAP) loading matrix [16].

To conduct JFA modeling it is necessary to estimate the speaker independent hyperparameters $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{D}$, $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$. In our work, we employ the coupled expectation-maximization (EM) algorithm hyperparameter training proposed by Vogt et al. [15].

2.2. CLR model comparison

After JFA modeling of the initial speaker segments, a robust metric is required to perform a pairwise comparison of the speaker models prior to clustering. We use the CLR metric as it has been shown to be a robust measure of pairwise similarity between models adapted using a UBM [10]. To do this, given two speaker segments i and j, and their corresponding feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively, the CLR score a_{ij} is computed as,

$$a_{ij} = \frac{1}{K_i} log \frac{p(\boldsymbol{x}_i|M_j)}{p(\boldsymbol{x}_i|M_B)} + \frac{1}{K_j} log \frac{p(\boldsymbol{x}_j|M_i)}{p(\boldsymbol{x}_j|M_B)}, \tag{2}$$

where, K_i and K_j represent the number of observations in $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively. M_i and M_j are the adapted models, and $p(\boldsymbol{x}|M)$ is the likelihood of $\boldsymbol{x}$, given model M, with M_B representing the GMM-UBM.

We then use the work by Glembek et al. [17], to accommodate CLR scoring into the JFA framework, calculating the likelihood function of model M, given data $\boldsymbol{x}$, using,

$$logp(\boldsymbol{x}|M) = \boldsymbol{Z}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{F} + \frac{1}{2} \boldsymbol{Z}^* \boldsymbol{N} \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}\,, \tag{3}$$

where, $\boldsymbol{\Sigma}$ is a $CP \times CP$ diagonal covariance matrix containing c, GMM components' diagonal covariance matrices, $\boldsymbol{\Sigma}_c$ of dimension $P \times P$. $\boldsymbol{N}$ is a $CP \times CP$ dimensional diagonal matrix consisting of each component's zeroth order Baum-Welch statistics, and $\boldsymbol{F}$ is a $CP \times 1$ dimensional vector achieved by concatenating the first order Baum-Welch statistics of each component. In our work, $\boldsymbol{F}$ was centralised on the GMM-UBM (M_B) mean mixture components.

2.3. Complete-linkage clustering

In our previous work we have demonstrated the efficiency and robustness of complete-linkage clustering [5], and have shown that this clustering method outperforms the traditional agglomerative cluster merging and retraining approach that is extensively used in speaker diarization [11, 18, 12, 13], as well as the alternative technique proposed by van Leeuwen [6], for carrying out agglomerative speaker clustering without retraining.

Complete-linkage clustering is a form of hierarchical clustering, in which the pairwise distance between clusters is employed to construct a clustering tree that represents the relationship between all speakers/clusters. The obtained tree can then be employed to merge clusters based on the complete-linkage criterion, and the final clustering outcome is then acquired using a distance threshold or the desired number of clusters [19].

In complete-linkage clustering models are initially merged based on a highest similarity, or lowest distance, score. As this clustering technique does not conduct retraining after each cluster merge, the pairwise scores between clusters are updated after a merge to indicate the distance between their most dissimilar elements. This approach thus takes into account the *best worst-case scenario* scores and assesses the relationship between all elements within two compared clusters, allowing for a more robust clustering decision.

To carry out complete-linkage clustering we first obtain the upper-triangular matrix $\boldsymbol{A}$, known as the attribution matrix [5], containing the pairwise CLR scores a_{ij} between all compared speaker models. As complete-linkage clustering is designed to compare distance values, as with our previous work [7, 5], from $\boldsymbol{A}$ we first compute an upper-triangular matrix $\boldsymbol{L}$, containing the corresponding pairwise distance scores l_{ij}, computed from the a_{ij} CLR scores using,

$$l_{ij} = \begin{cases} e^{(-a_{ij})}, & (i \neq j), \\ 0, & (i = j). \end{cases} \tag{4}$$

We then perform complete-linkage clustering using the distance attribution matrix $\boldsymbol{L}$, in the following manner:

1. Initialize $C{=}N$ clusters, assigning segment i to C_i.
2. Find the minimum distance score, l_{ij} and its corresponding clusters C_i and C_j.
3. Merge segments i and j by merging C_i and C_j into $C_{i'} = \{C_i, C_j\}$, and removing rows and columns i and j from $\boldsymbol{L}$.
4. Obtain the new $(N-1){\times}(N-1)$ matrix $\boldsymbol{L}$, by computing the distance between newly formed cluster and remaining clusters using the complete-linkage rule:

$$l_{i'r} = max(l_{ir}, l_{jr}) \tag{5}$$

5. If the stopping criterion is satisfied stop clustering, else repeat from step 2.

3. The SAIVT-BNEWS dataset

As speaker attribution is a recent area of research, there is a lack of availability of *suitable* datasets for evaluating proposed speaker attribution technology. A suitable evaluation corpus is one that provides reference diarization labels for each recording in the dataset, with multiple occurrences of speaker identities across recordings. In addition, a speaker identity key is required to ensure that each speaker, within each recording, can be mapped to a unique identity across the entire set of recordings. For this reason, in our previous work [7, 5, 8], we employed the National Institute of Standards and Technology (NIST) SRE 2008 summed channel telephone conversation test corpus [20]. This telephone corpus provides a range of inter-session data and allows for the convenience of employing a two-speaker stopping threshold for the diarization of each recording [3, 4, 5].

In this work, we collected a set of publically available Australian broadcast news recordings from a media website providing up to 100 broadcast news videos per day. We used this data to create a suitable attribution evaluation dataset, referred to as the SAIVT-BNEWS corpus. We did this to allow for free access to the data by other researchers active in the field of speaker attribution. We first collected a subset of the broadcast news data. This subset contained 55 broadcast news videos, centered on the same news topic and its related events. We selected the videos in this manner to ensure that the dataset contains multiple occurrences of unique speaker identities across recordings. We then extracted the audio, from the broadcast news videos, and manually produced reference diarization labels for each recording. To then identify the unique speaker identities across the set of recordings, we utilised the information in the video to label speakers across the recordings, allowing for the evaluation of speaker attribution across this subset of 55 recordings.

The 55 recordings collected range from 47 seconds to 5 minutes and 47 seconds in length. Each recording contains a different number of unique speaker identities, ranging from 1 speaker to a maximum of 9 speakers per analysed recording. As the recordings are from the broadcast news domain, a wide range of channel variations are observed both within and between recordings. Using reference diarization labels, a total of 175 initial speaker homogeneous segments are obtained, which can be linked to a total of 92 unique speaker identities across the entire dataset, consisting of 64 male and 28 female speakers.

A large variety of speakers are present in this dataset, such as reporters, politicians, children, elderly people and more. The presence of music in some videos and overlapping speech from different speakers provides an excellent corpus for evaluating the performance of attribution technology, as well as the possibility of addressing other new challenges. To obtain the SAIVT-BNEWS dataset, and its corresponding reference labels, the last author of this paper may be contacted by email.

4. Evaluation and results

In our previous work, we proposed a full speaker attribution system for conducting robust and efficient attribution of large datasets containing two-speaker telephone conversation recordings [7, 5, 8]. In this section we propose and evaluate a robust and efficient attribution approach that is applicable to multiple recording domains, with an arbitrary number of speakers within each recording. We begin by employing our telephone-data attribution system [5], and modify the diarization module of this system to accommodate recordings with any number of speakers, rather than only two speakers assumed for telephone conversations. We evaluate this baseline diarization approach on the SAIVT-BNEWS dataset (detailed in Section 3) to measure the performance of our previously proposed telephone-data diarization scheme, and reveal its robustness on a significantly different audio domain. We then analyse the shortcomings of our baseline diarization system and propose a simple modification to significantly improve the performance of this module.

After speaker diarization of the data, speaker linking is required to complete the task of speaker attribution. In this section, we propose employing our telephone-data speaker linking module [5, 8], to complete our multi-domain attribution system. We then evaluate our proposed attribution approach across the broadcast news dataset to demonstrate our system's performance across this corpus.

We evaluate the speaker diarization systems using the standard diarization error rate (DER) metric, as defined by NIST [20]. To evaluate our proposed speaker attribution system, we employ our previously proposed attribution error rate (AER) metric [5, 8]. In the studies conducted by van Leeuwen [6], and Vaquero et al. [4], cluster purity and coverage are used for evaluating speaker linking and attribution. We previously employed these measures to evaluate our system [7], however it is necessary to employ an error metric that reflects diarization errors, as well as the speaker linking errors. We believe the AER is a more appropriate metric for evaluating the task of attribution. The AER can be described as an extension to the standard DER measure, from a single recording, to a collection of recordings. The AER thus represents the percentage of time that a speaker identity is misattributed within recordings, as well as across recordings. To compute the AER it is necessary to first concatenate the reference diarization labels into a single label file and to then ensure that each unique speaker identity is labeled using a unique label across the entire concatenated reference label file. This can be referred to as the attribution reference label. The same process is then required to generate the attribution system label file, but this time based on the system's decision of the diarization output and the linked speaker identities. The two label files can then be compared using the NIST DER metric [20], however as this measured error is now representative of the DER per recording, as well as the speaker errors across recordings, we refer to it as the AER.

For JFA modeling the speaker and session subspaces were obtained using a coupled EM algorithm, with a 50-dimensional session and 200-dimensional speaker subspace [15]. The features we employed for speaker modeling were 13 MFCCs with 0^{th} order coefficient, deltas and feature warping [21], extracted using a 20 bin Mel-filterbank, 32 ms Hamming window and a 10 ms window shift. For the segmentation stages of our diarization module, as will be detailed in this section, we use 20 MFCCs with 0^{th} order coefficient, no deltas or feature warping, extracted in a similar manner. It is important to note that for JFA modeling of speaker segments, in both the diarization and speaker linking modules, we employ a previously trained combined gender GMM-UBM, consisting of 512 mixture components, trained using telephone speech data, as detailed in our previous work [7]. This means that our modeling approach is expected to perform better when dealing with telephone domain data. This work thus reveals the robustness of our attribution approach with respect to processing of multi-domain data.

4.1. Speaker diarization

As our baseline diarization system, we employ our previously proposed telephone-data speaker diarization module [5]. This

system was designed to perform robust and efficient diarization of two-speaker telephone conversation recordings. In this system, we followed the common practice of telephone-data diarization [4, 3], and employed our prior knowledge of the number of speakers within each recording as the stopping criterion to the clustering stage of our diarization module. We now require a method of dealing with an arbitrary number of speakers. Recall from Section 2.3, complete-linkage clustering can be carried out using the desired number of output clusters, or a distance threshold, as the stopping criterion to the clustering process. As we have no prior knowledge of the number of speakers within each recording, we propose using a suitable CLR threshold as the stopping criterion to the clustering phase of diarization. We thus go back to the CLR computation in (2),

$$a_{ij} = \overbrace{\frac{1}{K_i} log \frac{p(\boldsymbol{x}_i|M_j)}{p(\boldsymbol{x}_i|M_B)}}^{\delta_i} + \overbrace{\frac{1}{K_j} log \frac{p(\boldsymbol{x}_j|M_i)}{p(\boldsymbol{x}_j|M_B)}}^{\delta_j}, \qquad (6)$$

where (6) displays two splits of the CLR measure, δ_i and δ_j. δ_i represents likelihood that the data for speaker i is produced by the competing speaker model M_j, compared to the likelihood of this data being produced by the general speaker population (GMM-UBM). δ_j is the same measure, but for speaker j. From (6), a_{ij} will be negative if the general speaker population better models a speaker than its competing model, and a positive a_{ij} signifies that the speaker data in i and j are more similar to each other compared to the general speaker population. If ideal models are used, we would not expect δ_i and δ_j to have opposite signs and high absolute values, as it does not make sense for speaker i to be very similar to j but for j to be very different to speaker i. For these reasons, $a_{ij} = 0$ would serve as a suitable theoretical CLR threshold. We thus employ $a_{ij} \leq 0$ as the stopping criterion to the clustering stage of our diarization module to deal with an arbitrary number of speakers.

4.1.1. Baseline diarization system

We previously proposed a speaker diarization method using complete-linkage clustering for conducting efficient diarization within our proposed speaker attribution system [5] In this diarization system, we employ the hybrid voice activity detection (VAD) and the ergodic hidden Markov model (HMM) Viterbi resegmentation approach presented in [11]. We first use Viterbi segmentation to achieve an initial segmentation of the recordings, and then carry out modeling and clustering of these segments to complete the diarization process. We then apply a final Viterbi segmentation of the output speaker/clusters to refine the segment boundaries. In this work, we employ this system as our baseline diarization module and apply the CLR threshold stopping criterion, discussed in Section 4.1.

Our baseline system consists of the following stages:

1. Linear segmentation of the audio into 4 second segments and 3 iterations of Viterbi using 32 component GMMs to model each segment.
2. VAD to remove non-speech regions, followed by JFA modeling with session compensation.
3. Clustering of the speaker segment models using complete-linkage clustering until the CLR stopping threshold of $a_{ij} \leq 0$.
4. 3 iterations of Viterbi using 32 component GMMs to model final speaker/cluster, and a single Gaussian to model non-speech regions.

Table 1: *DER of baseline and proposed diarization systems.*

Diarization system	DER
Baseline	33.1%
Baseline + (1 iteration CLC)	**13.3%**
Baseline + (2 iterations CLC)	16.7%

4.1.2. Proposed diarization system and results

We evaluated our baseline diarization approach on the Australian broadcast news data, detailed in Section 3. The result of this evaluation can be seen in Table 1. It can be seen that our baseline diarization module is highly erroneous. We thus investigated the output of the baseline system to understand the underlying cause of the high DER obtained across the broadcast data. Through this investigation we found that our baseline system was under-clustering the speaker segments provided by the initial Viterbi segmentation and VAD stages. This may be addressed by knowing the desired number of output speakers, or by applying a different CLR stopping threshold (than 0) to the clustering process for each recording. However, this would mean having to abandon the convenience of employing a robust and theoretically ideal CLR threshold for any given recording. As our previous work on attribution [5], and particularly linking [8], had suggested that a CLR threshold value of 0 would serve as a robust stopping criterion, we concluded that the system was failing to robustly cluster speaker models as the initial segmentation did not provide sufficient data for modeled segments.

To overcome this, we propose using an additional pass of the complete-linkage clustering stage followed by Viterbi refinement. For convenience, we call the combination of these stages (steps 3 and 4 from Section 4.1) CLC, for complete-linkage clustering. We thus utilise the full baseline system to conduct a reliable initial segmentation of the recording, producing larger speaker homogeneous segments of data. We then apply a single iteration of CLC to the output of the baseline system. From Table 1 it can be seen that an absolute improvement of almost 20% is observed with respect to the DER measure.

This motivated our evaluation of another diarization system using the baseline system plus two additional passes of CLC. This system displayed a higher error rate than our proposed system using only one additional iteration of CLC. After observing the results, we found that a second additional iteration of CLC did not over-cluster the results, but it was rather the extra Viterbi refinement iterations that led to a higher DER measure, which reinforces our choice of the CLR stopping criterion of $a_{ij} \leq 0$. We thus propose employing our (baseline + CLC) diarization module for conducting robust speaker attribution.

4.2. Speaker attribution

In this section we employ our diarization system proposed in Section 4.1. As our previously proposed speaker linking system using complete-linkage clustering [5, 8], can be applied to this task without further modifications, we employ this linking module together with our proposed diarization method to carry out speaker attribution of the broadcast news data.

To conduct attribution, our proposed linking system obtains an initial set of (ideally) speaker homogeneous segments from the output of the diarization module across the collection of recordings. Each segment represents a unique speaker identity within its associated recording. These segments are then mod-

eled using JFA with session compensation, compared using the CLR metric and clustered using complete-linkage clustering.

We carried out the speaker attribution of the SAIVT-BNEWS data using our proposed multi-domain attribution system, which we will refer to as the **D-L** system, for diarization and linking. For evaluation purposes, we also carried out speaker attribution using reference diarization labels (DER = 0%) to initialise the speaker segment models in the linking phase of attribution. We did this for evaluation purposes and to reveal the potential of our attribution approach, should an ideal diarization module is used. To distinguish this system from our attribution approach, we will refer to this system as the **REF-L** system, for reference diarization and linking.

Figure 1 displays the AER of each system at all possible CLR threshold values. The horizontal axis has been reversed to display, from left to right, the clustering of the initial speakers/clusters into a single cluster. The oracle AER point of each system, obtained at its corresponding CLR threshold, has been marked on both the **D-L** and **REF-L** plots. It can be seen that as more speakers are correctly clustered a low AER region appears in the performance plot of each system. A lower valley, with respect to the vertical axis, indicates a higher accuracy associated with the analysed attribution system. In addition, the robustness of the systems is directly proportional to the width of the low AER region, and inversely proportional to the absolute value of the slope to the right of the oracle AER point, as marked on each plot. This slope is formed as each attribution system achieves its oracle AER point and then begins to attribute incorrect speaker identities to the already obtained clusters, creating a rise in the AER measure until all speakers are merged into a single cluster and maximum AER of the system achieved.

Table 2 displays the details associated with the oracle AER point of the two attribution systems. For reference, 92 unique speakers are present in the dataset, as detailed in Section 3. It can be seen that, as expected, the **REF-L** performs better than the **D-L** attribution system. This is also the case in Figure 1, which demonstrates that the **REF-L** system consistently performs better than the **D-L** attribution system. In addition, the CLR thresholds at which the oracle AER points of the two systems are achieved are both close to 0, thus further reinforcing the robustness of this CLR threshold as a stopping criterion to the task of clustering.

From Figure 1 and Table 2, it can be seen that the difference in the oracle AER of the two systems is almost equal to the DER displayed by our diarization module (Section 4.1). As the AER metric measures both the DER and the linking errors, and the fact that this difference in the oracle AER points of the two systems is almost equal to our achieved DER across the data, and as both systems achieve the same number of unique speaker identities across the dataset, it can be concluded that our linking module has been robust enough to deal with the erroneous diarization output. This suggests that any improvements to the DER achieved by our proposed diarization approach will directly apply to the AER obtained by our **D-L** system, potentially achieving a minimum AER of 17%, as obtained by our **REF-L** attribution system.

5. Discussion

Compared to our previous work on attribution of two speaker telephone-data [7, 5, 8], our multi-domain speaker attribution system proposed in this paper demonstrates similar results across the Australian broadcast news dataset. This is while our system remains largely unchanged, with the exception of the modification applied to the diarization module (Section 4.1) to accommodate an arbitrary number of speakers. Most importantly, as discussed in Section 4 and detailed in our previous work [7], our proposed multi-domain system employs a 512 component combined gender GMM-UBM, trained on telephone-data, for JFA modeling. This is indicative of the robustness of our attribution approach and suggests that our system may be improved even further through utilising a GMM-UBM trained on data from a broadcast news domain.

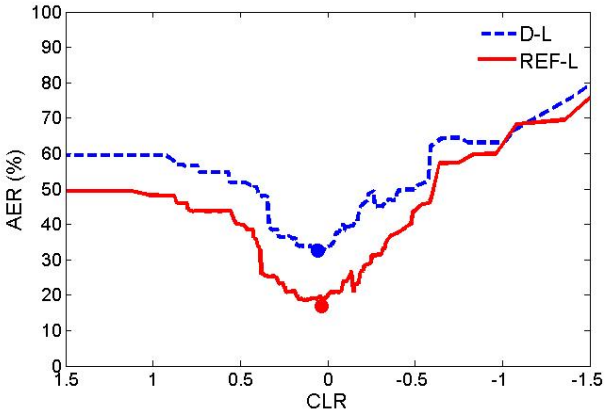

Figure 1: AER versus CLR for **REF-L** and **D-L** attribution.

Table 2: *Oracle attribution using* **REF-L** *and* **D-L** *systems.*

Attribution system	AER	Obtained speakers	CLR
REF-L	**17.0%**	77	0.03
D-L	32.6%	77	0.05

6. Conclusion

In this paper we proposed a robust and efficient speaker attribution approach, applicable to multiple audio domains, with the ability to conduct automatic diarization and attribution of multiple recordings, each containing speech from an arbitrary number of speakers. We did this by extending our previously proposed telephone-data speaker attribution approach. In this work, we proposed using a theoretically suitable CLR stopping threshold for complete-linkage clustering in diarization and linking. We demonstrated that, even in diarization where small segments are required to be clustered, this stopping threshold can be employed as a robust stopping criterion. Our work in this paper, and previous studies, suggests that this stopping threshold is robust across different audio domains when employed in the same manner as our multi-domain attribution approach. Finally, we demonstrated achievable AERs as low as 17%, across the broadcast news data, using our attribution system.

7. Acknowledgments

This paper was based on research conducted through the Australian Research Council (ARC) Linkage Grant No: LP0991238 and the follow-up applied research based on Australian broadcast data conducted through the Cooperative Research Centre for Smart Services.

8. References

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

[2] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *INTERSPEECH*, 2009, pp. 1559–1562.

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

[4] C. Vaquero, A. Ortega, and E. Lleida, "Partitioning of two-speaker conversation datasets," in *Interspeech 2011*, August 28-31 2011, pp. 385–388.

[5] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, march 2012, pp. 4185 –4188.

[6] D. A. V. Leeuwen, "Speaker linking in large data sets," in *Odyssey2010, the Speaker Language and Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 202–208.

[7] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Extending the task of diarization to speaker attribution," in *Interspeech2011*, Florence, Italy, August 2011. [Online]. Available: http://eprints.qut.edu.au/43351/

[8] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker linking using complete-linkage clustering," in *to be presented in Australian International Conference on Speech Science and Technology (SST2012)*, 2012.

[9] M. Ferras and H. Bourlard, "Speaker diarization and linking of large corpora," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, Dec., pp. 280–285.

[10] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1505 –1512, 2006.

[11] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. Springer Berlin / Heidelberg, 2008.

[12] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, nov.-3 dec. 2003, pp. 411 – 416.

[13] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557 –1565, 2006.

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

[15] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech 2008*, 2008, pp. 853–856.

[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[17] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 0, pp. 4057–4060, 2009.

[18] X. Anguera, S. Bozonnet, N. W. D. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech & Language Processing*, pp. 356–370, 2012.

[19] A. Jain, A. Topchy, M. Law, and J. Buhmann, "Landscape of clustering algorithms," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1, 2004, pp. 260 – 263 Vol.1.

[20] (2007) The NIST rich transcription website. http://www.nist.gov/speech/tests/rt/.

[21] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, June 18-22 2001, pp. 213–218.