A SPEAKER REDIARIZATION SCHEME FOR IMPROVING DIARIZATION IN LARGE TWO-SPEAKER TELEPHONE DATASETS

Houman Ghaemmaghami, David Dean, Sridha Sridharan

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

ABSTRACT

In this paper we propose a novel scheme for carrying out speaker diarization in an iterative manner. We aim to show that the information obtained through the first pass of speaker diarization can be reused to refine and improve the original diarization results. We call this technique speaker rediarization and demonstrate the practical application of our rediarization algorithm using a large archive of two-speaker telephone conversation recordings. We use the NIST 2008 SRE summed telephone corpora for evaluating our speaker rediarization system. This corpus contains recurring speaker identities across independent recording sessions that need to be linked across the entire corpus. We show that our speaker rediarization scheme can take advantage of inter-session speaker information, linked in the initial diarization pass, to achieve a 30% relative improvement over the original diarization error rate (DER) after only two iterations of rediarization.

Index Terms— Speaker rediarization, diarization, speaker linking, complete-linkage clustering, cross-likelihood ratio

1. INTRODUCTION

The rapid expansion of spoken archives around the world has brought about the need for technology capable of automatically annotating large volumes of spoken recordings with the identities of the speakers present in the analysed archive. This can be achieved through determining *'Who spoke when?'* within a recording (referred to as speaker diarization) [1, 2], and then identifying recurring speaker identities between recordings. In the studies carried out by Viet et al. [3] and Yang et al. [4] this task has been referred to as cross-show speaker diarization. We believe that although speaker diarization is a necessary module for carrying out such a task, the term diarization does not accurately reflect the problems associated with extending diarization to a collection-wide approach. When we compare within-recording speaker identities across temporally-independent recordings we may be faced with challenges such as inconsistencies in the speakers' recording environments or changes in the voice of the speakers due to aging or health complications. The study by van Leeuwen on cross-recording speaker identification [5], refers to this task as speaker linking, which has since become a widely accepted term in the field [5–9]. We also use the term speaker linking to distinguish this task from speaker diarization and have used the term *speaker attribution* in this paper (and our previous work) to refer to the combined tasks of speaker diarization and speaker linking [6, 10].

In recent literature, speaker attribution has been used as a vital module for person recognition in multimodal conditions using broadcast video data [9, 11]. As the size of the analysed collection grows, so too does the demand for greater efficiency. One challenge that directly impacts the efficiency of attribution is the clustering method employed for speaker linking. In diarization, agglomerative model merging and retraining is commonly used for clustering within a recording [1, 2]. This approach has also been applied to speaker attribution [3, 4]. This retraining approach is inefficient and is not feasible for dealing with large datasets. As an alternative, van Leeuwen first proposed a linking system using a form of agglomerative distance clustering [5], to overcome the inefficiencies of model agglomeration. We proposed complete-linkage clustering, a form of agglomerative distance clustering, which eliminates the need for model retraining and utilises a pessimistic distance rule to update pairwise cluster scores after each merge [6, 10].

After performing speaker attribution over a corpus with recurring identities across recordings, we achieve a set of hypothesised annotations. These annotations show where we believe each speaker is speaking in a single recording and other independent recording/s in which an identity is speaking. These annotations are not ideal and may contain erroneous information. We propose, for the first time, a novel scheme for using this additional information that is linked between independent recordings by a real diarization system, to redo and refine the initial diarization process of each recording. We call this *speaker rediarization* and investigate the effect of iteratively applying rediarization to boost diarization accuracy across the NIST 2008 SRE telephone corpus. We propose using the cross-likelihood ratio (CLR) metric to predict improvement after rediarization for two-speaker telephone recordings, allowing for blind selection/rejection of new annotations.

This research was supported by an Australian Research Council (ARC) Linkage Grant (No: LP130100110).

2. SPEAKER MODELING AND CLUSTERING

We utilise a joint factor analysis (JFA) modeling approach with session compensation [12], based on the method proposed by Vogt et al. [13]. We adapt models using a combined gender universal background model (UBM) [12, 13]. We first introduce a constrained offset of the speaker-dependent Gaussian mixture model (GMM) mean supervector:

$$\mathbf{m}_i(s) = \mathbf{m}(s) + \mathbf{U}\mathbf{x}_i(s), \tag{1}$$

where $\mathbf{m}(s)$ is the speaker-dependent, session-independent, GMM mean supervector of dimension $CL \times 1$, C is the number of mixture components used in the GMM-UBM and L is the dimension of the feature vectors. $\mathbf{x}_i(s)$ is a low-dimensional representation of variability in session i, and $\mathbf{U}$ is a low-rank transformation matrix from the session subspace to the UBM supervector space. Then,

$$\mathbf{m}(s) = \mathbf{m} + \mathbf{V}\mathbf{y}(s) + \mathbf{D}\mathbf{z}(s), \tag{2}$$

where $\mathbf{m}$ is the speaker- and session-independent GMM-UBM mean supervector. $\mathbf{y}(s)$ is the speaker factors, which represent the speaker in a specified subspace with a standard normal distribution [13]. $\mathbf{V}$ is a low-rank transformation matrix from the speaker subspace to the GMM-UBM mean supervector space. $\mathbf{D}\mathbf{z}(s)$ is the residual variability not captured by the speaker subspace, where $\mathbf{z}(s)$ is a vector of hidden variables with a standard Gaussian distribution, $N(\mathbf{z}|\mathbf{0}, \mathbf{I})$. $\mathbf{D}$ is the diagonal relevance maximum *a posteriori* (MAP) loading matrix. We estimate the speaker independent hyperparameters $\mathbf{U}$, $\mathbf{V}$, $\mathbf{D}$, $\mathbf{m}$ and $\mathbf{\Sigma}$, using the coupled expectation-maximization (EM) algorithm hyperparameter training by Vogt et al. [13].

Once JFA models are obtained, we employ the CLR metric for a pairwise comparison. The CLR has been shown to be a robust measure of model similarity when comparing adapted models [1]. In addition, CLR scoring can be accommodated by JFA modeling.

Given two segments i and j, and their feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively, the CLR score a_{ij}, indicating similarity of the segments with respect to their identity, is computed as,

$$a_{ij} = \frac{1}{K_i} \log \frac{p(\mathbf{x}_i|M_j)}{p(\mathbf{x}_i|M_B)} + \frac{1}{K_j} \log \frac{p(\mathbf{x}_j|M_i)}{p(\mathbf{x}_j|M_B)}, \tag{3}$$

where, K_i and K_j represent the number of feature vectors in $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. M_i and M_j are the adapted models. $p(\mathbf{x}|M)$ is the likelihood of data $\mathbf{x}$, given the JFA model M, and M_B is the GMM-UBM representing general population.

We need to incorporate JFA into the CLR framework. This requires some notation to be defined. $\mathbf{\Sigma}$ is a $CL \times CL$ diagonal covariance matrix containing C, GMM components' diagonal covariance matrices, $\mathbf{\Sigma}_c$ of dimension $L \times L$. Using this, for a model M of c ($c = \{1, \ldots, C\}$) components, adapted using data in $\mathbf{x}$, the zeroth and first order Baum-Welch statistics can be obtained [12].

Now let $\mathbf{N}$ be a $CL \times CL$ dimensional diagonal matrix consisting of each component's zeroth order statistics diagonal block $\mathbf{N}_c$. Let $\mathbf{F}$ be a $CL \times 1$ dimensional vector achieved by concatenating the first order statistics $\mathbf{F}_c$ of each component. Using this, and the work by Kenny [12], the likelihood function providing the likelihood of segment $\mathbf{x}$ given a JFA model M, can be calculated using,

$$\log p(\mathbf{x}|M) = \mathbf{Z}^*\mathbf{\Sigma}^{-1}\mathbf{F} + \frac{1}{2}\mathbf{Z}^*\mathbf{N}\mathbf{\Sigma}^{-1}\mathbf{Z}\,, \tag{4}$$

where $\mathbf{N}$ and $\mathbf{F}$ of each segment were obtained over each component, c, of the GMM-UBM. In addition, $\mathbf{F}$ is centralised on the GMM-UBM (M_B) mean components, $\mathbf{m}_c$.

After the pairwise CLR scores are computed using (3) and (4), we employ complete-linkage clustering to obtain the final speakers/clusters. Complete-linkage clustering is a form of agglomerative clustering that employs a linkage rule to update the pairwise cluster scores after a merge [14]. For this reason, complete-linkage clustering can be carried out without a model retraining stage, using only the initial set of pairwise CLR scores. To do this, the most similar pair of clusters (with highest pairwise CLR score) are first merged to form a starting node. The pairwise score between this newly formed cluster and each remaining cluster is then updated to the CLR score between their most dissimilar elements. For example, if we begin by merging two clusters C_i and C_j into $C_{i'} = \{C_i, C_j\}$, the CLR score ($a_{i'x}$) between the newly formed cluster $C_{i'}$ and any remaining cluster C_x will be updated using the complete-linkage rule,

$$a_{i'x} = min(a_{ix}, a_{jx}). \tag{5}$$

This merge and update process is repeated until a stopping criterion is reached. When updating the pairwise scores, complete-linkage clustering takes into account the *worst-case scenario* scores to pessimistically reflect the outcome of each merge. We have previously shown that this clustering technique is more efficient and more accurate than the standard and state-of-the-art model merging and retraining techniques used for diarization or linking [6, 10].

3. SPEAKER ATTRIBUTION

We employ an improved version of our proposed speaker attribution system using complete-linkage clustering [10]. We present the diarization and linking modules of our system and evaluate this across the NIST SRE 2008 dataset [15].

3.1. Speaker diarization

Our proposed diarization system is inspired by the ICSI RT-07 diarization system by Wooters et al. [2], and the baseline method by Kenny et al. [12]. We utilise an implementation of the hybrid voice activity detection (VAD) and ergodic HMM Viterbi segmentation technique from the ICSI

RT-07 system [2]. For VAD and the segmentation stages of diarization, we use 20 MFCC features including the 0^{th} order coefficient, extracted using a 20 bin Mel-filterbank, 32 ms Hamming windows and a 10 ms window shift. For clustering we utilise 13 MFCC features including the 0^{th} order coefficient and deltas, with added feature warping [16]. We use a stopping criterion of 2 speakers, which is a common assumption for diarization of two-speaker telephone used by Patrick Kenny [12] and Vaquero et al. [17]. Our system consists of:

1. Hybrid VAD for 15 iterations, or until convergence.
2. Linear segmentation of audio into 3 second segments, modeling each segment using 32 component GMMs and 3 iterations of Viterbi segmentation.
3. Obtaining zeroth and first order Baum-Welch statistics for the N initial segments using the GMM-UBM.
4. JFA modeling and CLR scoring between all pairs of segments to achieve ($N \times N$) CLR score matrix.
5. Complete-linkage clustering of N segments.
6. Model non-speech as single Gaussian and each speaker with 32 component GMMs for 3 iterations of Viterbi.

3.2. Speaker linking

Our linking system uses complete-linkage clustering for efficiency [6, 10]. We use JFA modeling to overcome inter-session variability, with a previously trained combined gender GMM-UBM of 512 mixture components. We use a 50-dimensional session and 200-dimensional speaker subspace. Our speaker linking module takes the output of our diarization system to initialise the inter-recording speaker models and consists of the following stages:

1. Obtaining the zeroth and first order Baum-Welch statistics for the N initial speakers using the GMM-UBM.
2. JFA modeling and CLR scoring between all pairs of speakers to achieve ($N \times N$) attribution matrix [10].
3. Complete-linkage clustering of N segments.

3.3. Attribution results

We carry out attribution across the NIST SRE 2008 two-speaker summed telephone corpus [15]. We use this dataset as it contains reference diarization labels for each recording, as well as multiple occurrences of unique speaker identities across independent recordings with a global identity key for the participating speakers. This allows the mapping of identities within recordings, to identities across the corpus, making it suitable for evaluation of our system. The NIST SRE 2008 provides an identity key for 1382 speakers across 691 recordings, consisting of a total of 751 unique identities.

We evaluate our diarization module using the standard DER metric, as defined by the National Institute of Standards and Technology (NIST) [15]. To only evaluate linking of speakers between recordings, we employ reference diarization labels. The linking module can be treated as attribution using our linking system after ideal diarization. We thus employ our proposed attribution error rate (AER) metric to evaluate linking, as well as attribution [6, 10]. The AER is the same as the DER, obtained across the corpus when taking into account speaker errors for linking inter-recording identities. We have thus renamed it to distinguish between within-recording errors (DER) and, within- and between-recording errors (AER). The AER will always be larger than the DER as it displays this error as well as the between-recording errors. Table 1 presents the performance of our proposed speaker attribution system, with respect to the AER metric, across the evaluation corpus. The performance of the speaker linking module is also shown in this table. As previously mentioned, the linking module was evaluated using reference diarization labels. Our work in [10] and [6] provide further details of the performance of these systems. Our objective is thus to improve upon these results through speaker rediarization.

System	DER	AER	Speakers found
diarization + linking	7.34%	26.08%	798
linking	reference	18.68%	901

Table 1. Speaker attribution and linking performance before applying speaker rediarization.

4. PROPOSED SPEAKER REDIARIZATION

In our speaker attribution system, the diarization stage provides speaker linking with a large set of intra-recording speaker identities. The linking module then models these speakers and clusters them to provide links between speaker models, hypothesised to be the same identity across the corpus. The most obvious approach to rediarization is then to use this knowledge to find additional models to each intra-recording identity and use the additional information to refine the original diarization output. From the results in Table 1, it can be seen that, even when reference diarization labels are used, the speaker linking module can be erroneous (AER = 18.68%). We thus employ a different approach to our linking system, for selecting similar speaker models to each speaker in a recording and conducting speaker rediarization.

4.1. Selecting similar models

After diarization is carried out and the CLR measure is computed between all intra-recording speaker identities, we seek any additional models that may be of use for conducting speaker rediarization. We propose using any inter-recording speaker model that is *similar* to the speakers in each recording for rediarization. We define this similarity based on the value of the pairwise CLR metric (a_{ik}) computed between

inter-session speakers i and k using (3) and a CLR threshold, θ. We propose the following rediarization strategy. First, in each recording, for each speaker i and j, find set S of similar speakers that are dissimilar to the competing speaker in the same recording. We define the similarity set for speaker i as:

$$S_i = \{k|(k \neq i) \wedge (a_{ik} \geq \theta) \wedge (a_{jk} < \theta)\} \quad (6)$$

where $k = \{1, \ldots, N\}$ is the set of speaker labels for the N initial speakers obtained across the entire corpus. We then use the initial diarization labels for segmentation, treating speaker turns as the beginning of new segments, such that no two segments are attributed to one another. This allows for reversing of erroneous intra-recording clustering decisions that took place in the initial diarization pass. We then model each segment using 32 component GMMs for 3 iterations of Viterbi to refine speaker change points.

After obtaining C ideally homogeneous segments in a recording, JFA modeling and CLR scoring are carried out to obtain a $(C \times C)$ CLR score matrix. At this stage we allow the larger linked models in sets S_i and S_j to participate in the rediarization process. Hence achieving a $(M \times M)$ CLR score matrix, where $M = (C + |S_i| + |S_j|)$, and $|S|$ denotes number of speakers in similarity set S. We then cluster the $(M \times M)$ score matrix, using complete-linkage clustering, to obtain a $(M \times 1)$ vector $\mathbf{p}$, containing the final speaker labels. We use $\mathbf{p}$ to appoint labels i or j to each of the M participating segments, discarding the inter-recording models added to assist with rediarization. Finally, we apply 3 iterations of Viterbi resegmentation to refine the final boundaries.

We aim to use larger speaker models, linked across independent recordings, to boost our knowledge of intra-recording speakers and improve diarization of a single recording through guiding the clustering of smaller intra-recording segments. The larger participating inter-recording models would ideally serve as initial nodes to which the smaller segments can be attributed when performing rediarization.

4.2. Blind selection of improved diarization labels

Although we hope to improve accuracy in this manner, we cannot guarantee the linking of matching identities. As CLR is a pairwise similarity metric (higher is more similar), we propose obtaining this measure between the final speaker identities and interpreting a reduction in this measure as an indication of improvement. This is simple for two-speaker telephone data. We accept the labels obtained through rediarization if $(a'_{ij} < a_{ij})$ is satisfied, where a_{ij} represents CLR between the speakers before rediarization, and a'_{ij} is this score after rediarization.

4.3. Rediarization and attribution results

We employ a range of CLR thresholds (θ) to select similar models for rediarization. We apply 3 iterations of our rediarization scheme after diarization. Figure 1 provides the DER

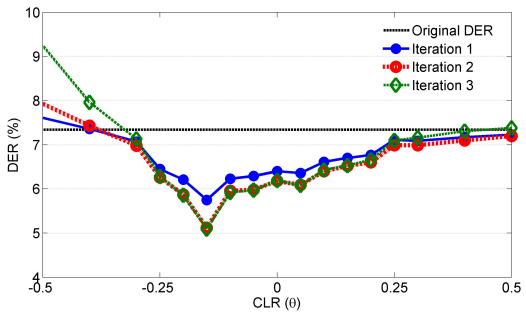

Fig. 1. DER at each rediarization iteration and CLR threshold.

Initial DER	Iteration 1	Iteration 2	Iteration 3
7.34%	5.75%	5.12%	5.11%

Table 2. DER decreases after applying each iteration of speaker rediarization at CLR threshold $\theta = -0.15$.

of our diarization module after each iteration of rediarization, at various CLR thresholds. Table 2 provides the performance of our rediarization system at its best performing threshold.

After two iterations of rediarization the DER is reduced from 7.34% to 5.12%. This is a relative improvement of 30.24% in DER. From Figure 1 and Table 2, virtually no improvement is observed after iteration 3 compared to iteration 2. This is because after iteration 2, the purity of the speaker models is not improved enough to reliably find *similar* models. The improvements are achieved within the CLR range of $\theta = [-0.25, 0.25]$. This can be explained using (3) and (7):

$$a_{ij} = \overbrace{\frac{1}{K_i} log \frac{p(\mathbf{x}_i|M_j)}{p(\mathbf{x}_i|M_B)}}^{T_i} + \overbrace{\frac{1}{K_j} log \frac{p(\mathbf{x}_j|M_i)}{p(\mathbf{x}_j|M_B)}}^{T_j}, \quad (7)$$

where T_i is the likelihood that data for speaker i is produced by the competing speaker model M_j, compared to likelihood of this segment being produced by the general population model (UBM). T_j represents this measure for speaker j. Hence, a_{ij} will be negative if the general population is a better representative of a speaker than its competing model. A positive a_{ij} signifies that speakers i and j are more similar to each other compared to the UBM. Given ideal models, we would not expect T_i and T_j to have opposite signs as we do not expect speaker i to be similar to j but for j to be different to speaker i. Hence, a CLR threshold of $(\theta = 0)$ would be an ideal threshold for choosing similar models. As larger thresholds are employed, the chance of correctly linking speakers is higher, however less recordings are found that meet the similarity constraint defined in (6). As θ is lowered more models are found, however once θ becomes too low $(\theta < -0.25)$ more false models are linked that lead to an increase in DER.

We found that our proposed blind label selection criterion was able to perform with great accuracy. Table 3 provides

Metric	Iteration 1	Iteration 2	Iteration 3
MR	1.2%	2.9%	3.6%
FAR	13.9%	18.4%	19.5%

Table 3. False alarm rate (FAR) and miss rate (MR) of blind selection of improved labels.

System	DER	AER	Speakers found
attribution	7.34%	26.08%	798
attribution + rediarization	5.12%	23.55%	897

Table 4. Speaker attribution evaluation results before and after applying rediarization show reduction in both DER and AER.

the performance of this criterion in terms of the miss rate (MR) and false alarm rate (FAR) of the criterion in successfully predicting improvement to the diarization labels. Table 4 displays the performance of our attribution system versus the performance of this system with rediarization. The AER of the attribution system is reduced through rediarization. This reduction is more than the absolute reduction in the DER metric, indicating that the improvement in DER has also provided the linking module with more pure models.

5. CONCLUSION

We proposed a novel speaker rediarization algorithm for improving the performance of speaker diarization and linking in large two-speaker telephone datasets. In this work we demonstrated that the information obtained through speaker diarization and linking across a corpus (although erroneous) can be used, through our proposed rediarization technique, to improve and refine the initial diarization outcome. We proposed an iterative rediarization system and evaluated this approach over the NIST SRE 2008 telephone corpus, showing a relative improvement of 30% in DER and 10% in AER. In addition, we proposed a blind label selection criteria, based on the CLR metric, for successfully predicting improvements to the original diarization labels after rediarization.

REFERENCES

[1] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE ASLP*, vol. 14, no. 5, pp. 1505 –1512, 2006.

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

[3] T. Viet-Anh, L. Viet Bac, C. Barras, and L. Lamel, "Comparing multi-stage approaches for cross-show speaker diarization.," in *INTERSPEECH*. 2011, pp. 1053–1056, ISCA.

[4] Q. Yang, Q. Jin, and T. Schultz, "Investigation of cross-show speaker diarization.," in *INTERSPEECH*. 2011, pp. 2925–2928, ISCA.

[5] D. A. Van Leeuwen, "Speaker linking in large data sets," in *Odyssey2010*, Brno, Czech Republic, June 2010, pp. 202–208.

[6] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker linking using complete-linkage clustering," in *SST2012*, 2012.

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

[8] M. Ferras and H. Bourlard, "Speaker diarization and linking of large corpora," in *IEEE SLT Workshop 2012*, Dec., pp. 280–285.

[9] H. Bourlard, M. Ferras, N. Pappas, A. Popescu-Belis, S. Renals, F. McInnes, P. Bell, S. Ingram, and M. Guillemot, "Processing and linking audio events in large multimedia archives: The eu inevent project," in *Proceedings of SLAM 2013*, 2013.

[10] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *IEEE ICASSP2012*, march 2012, pp. 4185 –4188.

[11] A. Giraudel, M. Carr, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus : a multimodal corpus for person recognition," in *Proceedings of LREC'12*, Istanbul, Turkey, may 2012.

[12] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal STSP*, vol. 4, no. 6, pp. 1059 –1070, 2010.

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

[14] A.K. Jain, A. Topchy, M.H.C. Law, and J.M. Buhmann, "Landscape of clustering algorithms," in *Proceedings of ICPR2004*, 2004, vol. 1, pp. 260 – 263 Vol.1.

[15] "The NIST year 2008 speaker recognition evaluation plan," Tech. Rep., NIST, 2008.

[16] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Odyssey2001*, June 18-22 2001, pp. 213–218.

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