SPEAKER ATTRIBUTION OF MULTIPLE TELEPHONE CONVERSATIONS USING A COMPLETE-LINKAGE CLUSTERING APPROACH

Houman Ghaemmaghami, David Dean, Robbie Vogt, Sridha Sridharan

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

ABSTRACT

In this paper we propose and evaluate a speaker attribution system using a complete-linkage clustering method. Speaker attribution refers to the annotation of a collection of spoken audio based on speaker identities. This can be achieved using diarization and speaker linking. The main challenge associated with attribution is achieving computational efficiency when dealing with large audio archives. Traditional agglomerative clustering methods with model merging and retraining are not feasible for this purpose. This has motivated the use of linkage clustering methods without retraining. We first propose a diarization system using complete-linkage clustering and show that it outperforms traditional agglomerative and single-linkage clustering based diarization systems with a relative improvement of 40% and 68%, respectively. We then propose a complete-linkage speaker linking system to achieve attribution and demonstrate a 26% relative improvement in attribution error rate (AER) over the single-linkage speaker linking approach.

Index Terms— speaker attribution, diarization, linking, agglomerative clustering, complete-linkage

1. INTRODUCTION

The task of *speaker attribution* refers to the annotation of a collection of spoken audio based on speaker identities [1]. In order to conduct attibution of multiple recordings, diarization must first be performed to obtain a set of speaker-homogeneous audio segments from each recording. This is followed by *speaker linking*, which aims to group segments from the same speaker across the collection of audio segments [2]. In order to employ speaker attribution for annotation of large audio archives, it is important to conduct this task in a robust and efficient manner [1]. To do this we must achieve efficiency in both the diarization and linking phases. Various robust diarization systems have been proposed but most rely on the traditional agglomerative merging and retraining approach to clustering [3, 4]. In recent work on speaker linking this approach has been simplified to single-linkage clustering through elimination of the retraining stage [2, 5, 6]. The issue with such methods is that traditional agglomerative systems are inefficient and single-linkage, or *nearest neighbour* clustering methods are associated with the *chaining effect*. This is the erroneous chaining of outer samples belonging to distinct clusters based on a minimum distance requirement, which promotes growth of elongated clusters [7].

This paper proposes an attribution system using a complete-linkage clustering approach. The proposed system extends our work in [1] and utilises a complete-linkage technique to conduct diarization and speaker linking. We employ joint factor analysis (JFA) using a combined-gender universal background model (UBM) to model clusters in the diarization and linking tasks [8]. The normalized cross-likelihood ratio (NCLR) is computed as the pairwise cluster similarity metric [4]. Finally, complete-linkage clustering is used to obtain the final clusters without retraining.

Section 2 provides the theory behind single- and complete-linkage clustering of JFA adapted speaker models. In Section 3 the linkage clustering diarization systems, and the agglomerative diarization method with merging and retraining are presented and evaluated. Section 4 describes the full speaker attribution system using single- and complete-linkage clustering based speaker linking. Section 4 also introduces the attribution error rate (AER) as an evaluation metric based on the diarization error rate (DER) and presents evaluation of the proposed attribution systems. Section 5 then concludes the paper with discussion of results and future work.

2. LINKAGE CLUSTERING OF SPEAKER MODELS

Linkage clustering is a form of agglomerative hierarchical clustering in which clusters are merged based on a rule set by the form of linkage clustering, and using a pairwise cluster “distance” metric [7]. We use a pairwise distance based upon the NCLR. The NCLR is not strictly a distance, it is in fact a measure of similarity as it does not obey the triangular equality, however it is convenient to refer to it as a distance.

The main difference between linkage clustering and traditional agglomerative methods is that linkage clustering does not involve merging and retraining of cluster models at each level of clustering. In linkage clustering the pairwise cluster distances can be used to form a tree representation of the relationship between the clusters. The final stage of clustering can then be achieved using a distance threshold or by choosing the required number of clusters as the stopping criterion.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

Our work is motivated by the simplicity and efficiency of linkage clustering. Throughout this paper, we employ the same approach for cluster modeling and computing the pairwise cluster distances in both diarization and speaker linking stages of attribution. To model clusters, we utilise a JFA approach and adapt models from a UBM, as described in [9]. After obtaining the cluster models we compute the normalized cross-likelihood ratio (NCLR) between pairs of adapted models. The NCLR is considered to be a robust similarity metric for adapted speaker models [10], and is computed as:

$$s_{ij} = \frac{1}{N_i} \log \frac{p(\boldsymbol{x}_i|M_j)}{p(\boldsymbol{x}_i|M_B)} + \frac{1}{N_j} \log \frac{p(\boldsymbol{x}_j|M_i)}{p(\boldsymbol{x}_j|M_B)} \qquad (1)$$

where, s_{ij} is the NCLR metric between clusters i and j, N_i and N_j are the number of observations for each cluster model M_i and M_j, respectively. $p(\boldsymbol{x}|M)$ denotes the likelihood of the data $\boldsymbol{x}$ given model M, and M_B represents the UBM. In [10], it is shown that for JFA adapted models:

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

where $\boldsymbol{N}$ and $\boldsymbol{F}$ represent 0^{th} and 1^{st} order statistics of the cluster segment $\boldsymbol{x}$ calculated using model M. $\boldsymbol{Z}$ is the sum of the speaker/cluster and channel supervectors. $\boldsymbol{\Sigma}$ is the covariance of the UBM. $\boldsymbol{F}$ and $\boldsymbol{N}$ were obtained for components of the UBM and $\boldsymbol{F}$ was centralized on the UBM mean.

As previously mentioned, linkage clustering uses pairwise cluster distances to conduct hierarchical clustering, however the NCLR is a similarity metric. For this reason we must ensure that our metric conforms to the assumptions of linkage clustering regarding the pairwise "distances", namely their nonnegativity, reflexivity, dissimilarity and symmetry prior to clustering [7]. From (1) it can be seen that the NCLR metric is symmetric and a similarity metric, hence we utilise the procedure in (3) to obtain d_{ij} from the raw NCLR values, s_{ij}. This achieves dissimilarity, non-negativity and reflexivity while maintaining the integrity of the NCLR metric:

$$d_{ij} = \begin{cases} e^{(-s_{ij})}, & (i \neq j), \\ 0, & (i = j) \end{cases} \qquad (3)$$

Using (3) we can convert the NCLR values, s_{ij} in the square speaker similarity matrix $\mathbf{A}$, to obtain a square dissimilarity matrix $\mathbf{A}'$ containing the d_{ij} metrics. After this procedure the top triangle of $\mathbf{A}'$, containing the d_{ij} values, can be used for linkage clustering. In order to once again ensure the integrity of the original NCLR scores we are limited to linkage clustering methods that employ the d_{ij} values without modification. We used single- and complete-linkage clustering:

- Single-linkage: $L(a,b) = min(d_{ai}, d_{bj})$,
- Complete-linkage: $L(a,b) = max(d_{ai}, d_{bj})$.

Where, $L(a,b)$ is a d_{ij} value associated with the link between between two clusters a and b for $i \in a$ and $j \in b$. It can be seen that the single-linkage method employs the minimum distance for clustering. This is not appropriate in the case where the clusters are not separated well-enough for a threshold value to achieve clustering without encountering the *chaining effect*. This motivated our investigation of the complete-linkage clustering approach. The advantage of complete-linkage is that it discourages the growth of elongated clusters, however this method is prone to clustering outliers when using large threshold values [7].

In our work, pairs of speaker models are linked using the d_{ij} metric. A threshold value for L may be imposed as a stopping criterion to terminate clustering. Alternatively, clustering can be terminated at a desired number of clusters.

3. SPEAKER DIARIZATION

We propose a linkage based speaker diarization system using single- or complete-linkage clustering and compare this to a traditional agglomerative approach with retraining. The diarization systems were inspired by the ICSI RT-07 system in [3], and the baseline method in [8]. To do this, we employ the hybrid voice activity detection (VAD) and the ergodic HMM Viterbi resegmentation approach described in [3]. We utilise the Viterbi decoding to achieve initial segmentation of the recordings. For this purpose we do not apply the VAD decisions until after the segmentation. This is to ensure that we take advantage of silence regions for the initial segmentation. The VAD decisions are only applied prior to clustering, however the non-speech regions are introduced back after clustering for a final Viterbi resegmentation using the speaker models and a non-speech model to refine boundaries. Finally, as we are conducting diarization of telephone conversations we utilise a stopping criterion of 2 clusters/speakers, this is a common assumption, as used in both [5] and [8].

3.1. Proposed linkage diarization

We implemented two linkage clustering based diarization systems using the linkage methods described in Section 2. To do this we applied the following to each recording:

1. Linear segmentation into 4 second segments and 10 iterations of Viterbi using 32 component GMMs to model segments.
2. VAD to remove non-speech regions, followed by single-linkage or complete-linkage clustering to two clusters/speakers as described in Section 2.
3. 10 iterations of Viterbi using 32 component GMMs to model each of the two speakers based on the approach in [8]. In addition, non-speech regions are reintroduced and modeled using a single Gaussian prior to the Viterbi resegmentation.

3.2. Agglomerative diarization with retraining

To avoid confusion, we will hereon refer to this system as the *retraining* system, as the linkage methods are also agglomerative approaches. This system first conducts segmentation using the exact method in stage 1 of Section 3.1. In stage 2, the retraining system conducts cluster merging based on the minimum d_{ij} value. After each merge the system retrains new models using the cluster modeling in Section 2. This is done until only two clusters remain. The final stage follows the exact same approach as stage 3 in Section 3.1.

3.3. Diarization results

We evaluated the linkage and retraining systems using 691 excerpts from the summed channel NIST SRE 2008 test data, consisting of two speaker conversations between 751 unique speakers. Each recording was approximately 5 minutes of summed two-speaker telephone speech with double-talk regions also included. For the Viterbi segmentation we used 20 MFCC features including the 0^{th} order coefficient. For JFA modeling we used 13 MFCCs with 0^{th} order coefficient, deltas and feature warping. Hyperparameter training of the UBM and JFA subspaces was conducted as in [1].

Table 1 displays the diarization error rates (DER) for the linkage systems and the retraining method. It is seen that the complete-linkage system greatly outperforms single-linkage diarization. This is due to the defects associated with utilising the shortest distance which promotes cluster chaining. The complete-linkage system also outperforms the retraining system with a relative improvement of more than 40% with respect to the DER. Our complete-linkage approach is also superior in efficiency as clustering is conducted without retraining of models. In the case of speaker attribution we are mostly concerned with efficiency, however the results indicate that a higher accuracy may also be achieved.

4. SPEAKER ATTRIBUTION

We showed that our complete-linkage diarization significantly outperforms the single-linkage and retraining methods. In addition, the study in [5] demonstrates that a higher DER can notably degrade the attribution performance. We thus only utilise our complete-linkage diarization system to conduct speaker linking and report attribution results.

We carry out speaker linking based on our approach in [1] using a complete-linkage linking system as well as a single-linkage speaker linking approach and compare the two systems. This was done to assess the performance of the single-linkage method when dealing with large inter-session speaker models for linking as opposed to short utterance models in diarization. We did not implement an agglomerative retraining speaker linking system. This is because traditional agglomerative retraining methods are not efficient, especially when dealing with large numbers of speaker models [2].

Table 1. *DER for the proposed linkage diarization systems and the retraining diarization system*

Diarization system	Diarization error rate
Complete linkage	**10.75%**
Single linkage	33.20%
Retraining	**17.95%**

4.1. Attribution results and discussion

For evaluation, we introduce the attribution error rate (AER). AER is a direct extension of DER to multiple recordings where segments that are not correctly clustered in the linking stage lead to an increase in speaker errors. To compute AER, reference diarization labels are concatenated, marking reference speaker identities. This is compared to the system label, which is obtained by concatenating the diarization output labels and attributed identities. As attribution is considered diarization followed by speaker linking, it is important to use a metric that reflects the errors associated with both tasks.

The linking modules were initialised with 1382 speakers (691 files×2 speakers). We did not set a stopping criterion for linking. This is beyond the scope of this paper, however for attribution we evaluated the systems for a range of thresholds corresponding to all possible attribution outputs to obtain the minimum achievable AERs. Figure 1 and Table 2 show the obtained results for the three evaluated systems: (complete diarization + single linking), (complete diarization + complete linking) and (reference diarization + complete linking).

Figure 1 displays how the AER is affected for each system at various thresholds and as a result of erroneous diarization. The horizontal axis has been reversed to display, from left to right, the clustering of 1382 speakers into a single speaker identity. The initial gap between the reference system and the dashed plots represents the difference in DER of the applied diarization system. As more speakers are clustered correctly it can be seen that a valley region appears in the plot; the lower this valley, the higher the accuracy of the system. The width of the valley may be associated with the robustness of the system. Once a system reaches it's minimum error point it then begins to cluster incorrect speaker identities which gives quick rise to the AER until all speakers are clustered. We can assess each system based on the sharpness of the slope to the right of the minimum AER point. It can be seen that, the single-linkage approach fails due to incorrect speaker clustering as a result of the chaining effect. This is apparent from the sharp slope and the sudden rise in the single-linkage plot. The slope and shape of the plots for the two complete-linkage based systems are similar with an almost constant gap between the two plots before the miminmum AER point. This gap represents the DER and indicates that the errors introduced by our diarization system limit the achievable attribution accuracy using our complete-linkage speaker linking

Table 2. *Minimum AER for linking systems and number of speakers obtained (true number of speakers in dataset is 751)*

Speaker linking	AER	NCLR	Speakers
Complete linkage	**29.15%**	**-0.03**	**796**
Single linkage	39.10%	0.14	1062
Reference complete-linkage	18.68%	0.07	902

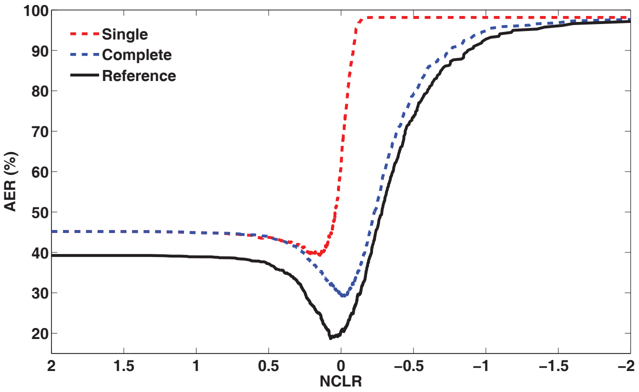

Fig. 1. AER versus NCLR for the three evaluated systems

system. Finally, from Table 2 it can be seen that the single-linkage system obtains the highest number of speakers at its minimum AER point. This is due to the chaining effect giving rise to the AER. In addition, the reference system achieves its minimum AER at a higher number of speakers than the complete-linkage based attribution. This may be due to the presence of short speaker segments that do not provide the sufficient amount of data required by our speaker modeling approach. These segments would not significantly impact the overall AER due to their short length but may be clustered independently, yielding a higher number of attributed speakers.

5. CONCLUSION

In this paper we proposed an attribution system based on complete-linkage clustering. We implemented a single- and a complete-linkage clustering diarization system as well as an agglomerative retraining system. Through evaluation of the diarization systems we demonstrated that the complete-linkage diarization approach outperformed both the single-linkage and agglomerative retraining systems with a relative improvement of 68% and 40% in DER over a subset of the NIST 2008 SRE corpus, respectively. In addition, we implemented a single- and a complete-linkage speaker linking system to conduct attribution using complete-linkage diarization. It was seen that the complete-linkage speaker linking outperformed the single-linkage speaker linking system with a relative improvement of 26% in AER over the test set. We then employed this linking approach to analyse the effects of erroneous diarization on the AER evaluation metric.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Extending the task of diarization to speaker attribution," in *Interspeech2011*, Florence, Italy, August 2011.

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

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

[4] O. Mella L. Viet Bac and D. Fohr, "Speaker diarization using normalized cross likelihood ratio," in *Interspeech 2007*, August 2007.

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

[6] N. Brummer and E. de Villiers, "The speaker partitioning problem," in *Odyssey2010, the Speaker Language and Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 194–201.

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

[8] 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.

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

[10] 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.