An Iterative Speaker Re-Diarization Scheme for Improving Speaker-Based Entity Extraction in Multimedia Archives

Houman Ghaemmaghami, David Dean, Sridha Sridharan

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

Abstract

In this paper we propose a practical speaker re-diarization scheme for improving speaker diarization by making use of repeating speakers across multiple recordings within a large corpus. We call this technique speaker re-diarization and demonstrate that it is possible to reuse the initial speaker-linked diarization outputs to boost diarization accuracy within individual recordings. We first propose and evaluate two novel re-diarization techniques. We demonstrate their complementary characteristics and fuse the two techniques to successfully conduct speaker re-diarization across the SAIVT-BNEWS corpus of Australian broadcast data. This corpus contains recurring speakers in various independent recordings that need to be linked across the dataset. We show that our speaker re-diarization approach can provide a relative improvement of 23% in diarization error rate (DER), over the original diarization results, as well as improve the estimated number of speakers and the cluster purity and coverage metrics.

Index Terms: speaker re-diarization, diarization, speaker linking, complete-linkage clustering, cross-likelihood ratio

1. Introduction

The rapid increase in multimedia archives around the world has drawn attention to technologies capable of automatically annotating large sets of recordings with respect to the speaker identities in the analysed set. Speaker diarization is a necessary module for conducting this task and can be utilised to reveal ‘Who spoke when?’ in a given recording [1]. Viet et al. [2] and Dupuy et al. [3] extended speaker diarization, from identifying speakers within recordings, to detecting recurring speaker identities across independent recordings. They refer to this process as cross-show speaker diarization. As speaker diarization is traditionally applied within a recording [4, 5], it is not necessarily designed to identify speakers across temporally independent recordings, which may contain inconsistencies in recording environments or the voice of speakers due to aging or health related complications. These variations are particularly apparent when analysing archives of broadcast television data, where a presenter (or a unique entity) may reappear across recordings separated by decades in time. For this reason, we employ the term speaker linking, as first used by van Leeuwen [6] and adopted by many others [7, 8, 9, 10], to refer to the task of determining speakers across temporally-independent recordings after diarization has been applied to extract speakers within each recording. For consistency with our previous work [11, 12, 13, 9], we use the term speaker attribution to refer to the combined tasks of speaker diarization and speaker linking.

In this paper we propose a practical speaker re-diarization scheme for improving speaker attribution performance across a dataset, without any prior knowledge of the data. For evaluations, we use the SAIVT-BNEWS corpus of Australian broadcast data [9]. To conduct speaker re-diarization, we first require an annotation of the SAIVT-BNEWS dataset as an initial hypothesis of the identities present in the corpus, the recordings they appear in and when they speak in each recording. We obtain this annotation using our previously proposed speaker attribution system [9], which we evaluate over the dataset to report a baseline performance. We then aim to improve the baseline performance using our proposed re-diarization technique. We first propose a cluster recombination scheme to re-diarization, for reincorporating the initial annotation obtained using the baseline. We show that this technique is biased toward boosting cluster coverage and thus propose a complementary purification scheme to increase purity. We then fuse the two approaches in a joint scheme to achieve a 23% relative improvement in diarization error rate (DER) over the baseline performance.

2. SAIVT-BNEWS evaluation corpus

We employ the SAIVT-BNEWS evaluation corpus [9], which is a publically available collection of Australian broadcast television data. This corpus contains 55 videos, most of which are news programs with inter-related topics that allow for recurring identities across multiple recordings and session conditions. The dataset is provided with a set of reference annotation labels that can be used for evaluation and contains a large variety of speakers, such as reporters, politicians, presenters, children and elderly people. It also contains overlapping speech segments and music, all of which we take into account when reporting evaluation metrics. The 55 files in the dataset range from 47 seconds to 5 minutes and 47 seconds, contain from 1 to a maximum of 9 unique speakers within each recording, with a total of 92 globally unique identities across the dataset.

3. Baseline speaker attribution

We employ our speaker attribution system proposed in [9], as our baseline system before applying re-diarization. We first provide a brief description of this system.

3.1. Speaker Modeling and Clustering

The baseline system employs joint factor analysis (JFA) modeling with session compensation [14, 15], which makes it ideal for modeling and comparing a variety of speakers across different session conditions. This is detailed in our previous work [13, 9]. We adapt models using a combined-gender UBM of 512 mixture components, with a 50-dimensional session and 200-dimensional speaker subspace. We train the speaker independent JFA hyperparameters using a coupled expectation-maximization (EM) algorithm proposed by Vogt et al. [15].
After the JFA adapted models are obtained for each participating speaker, a pairwise CLR similarity score is computed between the models to decide if they are to be merged. The CLR score has been shown to be a robust metric for comparing speaker models [4], especially when combined with complete-linkage clustering [12]. Based on our previous work [12, 9], employing the CLR metric in this manner appears to provide a natural comparison threshold value of 0.0 that, although not ideal, can be used as a stopping criterion to attribution across multiple audio domains with reasonable results. As we are performing speaker re-diarization with the aim of improving the initial results, we only require the baseline attribution system to provide an estimate of the speaker models and the location of their associated speech segments. We thus use a CLR threshold value of 0.0 as our stopping criterion to speaker clustering.

We employ complete-linkage clustering to obtain the final clusters. Complete-linkage is a form of agglomerative clustering that employs a rule, known as the linkage rule, to update pairwise scores after a merge [16]. We have previously shown that this approach outperforms traditional agglomerative clustering with retraining [5] and other state-of-the-art techniques [6] employed for clustering in speaker attribution [13, 12]. It can be carried out without a retraining stage, using only the initial CLR scores. This is done by first merging the most similar pair of clusters to form a starting node. The pairwise score between this new cluster and each of the remaining clusters is then updated to reflect the CLR score between their most dissimilar elements. For example, if we merge two clusters \(C_i\) and \(C_j\) into \(C_{ij} = \{C_i, C_j\}\), the score between the newly formed cluster \(C_{ij}\) and any remaining cluster \(C_k\) will be \(a_{ij,k}\), where,

\[
a_{ij,k} = \min(a_{i,k}, a_{j,k}).
\]

\[1\]

### 3.2. Baseline Diarization and Linking

We use our previously proposed diarization approach [9], as our baseline diarization system. Our system does not require tuning and has been shown to be robust across multiple domains (broadcast news and telephone) [9, 13]. This system is inspired by the ICSI RT-07 diarization system by Wooters et al. [5], and the baseline method by Kenny et al. [17]. We use an implementation of the hybrid voice activity detection (VAD) and ergodic HMM Viterbi segmentation from ICSI RT-07 [5]. For a given recording, we first conduct linear segmentation of the audio using 4 second segments, followed by 3 iterations of Viterbi using 32 component GMMs to model each segment. We then remove non-speech regions using hybrid VAD [13, 9]. This provides us with ideally homogeneous segments for modeling and clustering using Section 3.1. After clustering we perform 3 iterations of Viterbi using 32 component GMMs for each obtained speaker/cluster and a single Gaussian for non-speech. We then repeat the last two steps (clustering and Viterbi segmentation), as discussed in our previous work [9].

Our baseline speaker linking system is responsible for linking speaker identities, unique to each recording, across independent recordings within the corpus. This is carried out using the process detailed in Section 3.1. The system takes as input the diarization output labels (obtained using Section 3.2) and models each segment using JFA, measures similarity using CLR and clusters using complete-linkage clustering [12, 9].

### 3.3. Baseline Evaluations

We evaluated our baseline system over the SAIVT-BNEWS corpus presented in Section 2. We use the standard DER [18], cluster purity (CP) and cluster coverage (CC) [11], evaluation metrics throughout this paper. We first conduct diarization and then speaker linking to achieve attribution. We will refer to the diarization error, which reflects the within-recording errors, using DER. To distinguish between the diarization error and the error associated with speaker attribution (diarization and linking), we use the term attribution error rate (AER). AER is the DER computed within and between independent recordings, thus taking into account recurring identities across the corpus.

Table 1 displays the performance of our baseline system. The first row provides the diarization performance in DER, CP and CC, as well as the number of unique speakers found within each recording before speaker linking. The second row then shows the same metrics for our attribution system. To save space, we utilise the heading ‘ER%’ (for error rate) to indicate the DER and AER metrics throughout this paper. It can be seen that the errors associated with diarization are carried to the linking stage, resulting in a higher error rate (AER). It is also seen that after applying speaker linking to diarization, the number of speakers is reduced from 166 intra-recording speakers to 67 unique identities. This number is lower than the true 92 speakers in the dataset. This is mainly due to incorrect attribution of speakers with short utterance samples to other identities.

| Table 1: Baseline diarization and attribution evaluations. |
|-----------------|------|------|------|------|
| **Baseline**    | **ER%** | **CP%** | **CC%** | **Speakers** |
| **Diarization** | 13.2  | 80.8  | 92.6  | 166   |
| **Attribution** | 35.9  | 74.6  | 74.9  | 67    |

### 4. Speaker Re-Diarization

Although the baseline attribution system has errors, we believe it also provides useful information that we initially did not have. This is information for hypothesised identities, obtained across multiple recordings, that can be used as additional knowledge in each of the recordings that the identities appeared in. We aim to show that it may be possible to utilise this information to repeat diarization and reduce the initially obtained error metrics. We refer to this as speaker re-diarization, which extends our work on speaker re-diarization of two-speaker telephone data [19]. We first propose a cluster recombination approach for using the additional speaker information to reapply diarization. We then demonstrate the pitfalls of this technique and propose a purification scheme to overcome these issues in a joint cluster purification-recombination scheme.

#### 4.1. Re-diarization by Recombination

After conducting attribution, we obtain a set of speaker models and their annotation labels, which are hypothesised to be unique across the analysed set. As we use JFA modeling [17], we store the speaker models in the form of zeroth and first order Baum-Welch (BW) statistics for efficiency in future CLR scoring and clustering [9, 13]. We propose introducing each globally unique speaker model, as hypothesised by our baseline attribution, into the clustering stage of the diarization of its associated recordings, to perform re-diarization. We refer to our proposed scheme as cluster recombination, or recombination. We thus aim to enrich the within-recording diarization process with speaker segments that are not from that recording, but are linked to that recording by the baseline system, as they are hypothesised to belong to the same global speaker identity.

When conducting attribution in an archive \(R\) with \(N\)
recordings, such that $R = \{R_n; n = 1, \ldots, N\}$, we begin the diarization process by obtaining a segmentation for each recording $R_n$. This is a set $X_n = \{X_{n,k}; k = 1, \ldots, K_n\}$, containing BW statistics for $K_n$ adjacent, and ideally speaker-homogeneous, segments in that recording. These segments are not yet clustered to achieve diarization. We then carry out the clustering phase of diarization to obtain and store a set of BW statistics for locally unique speakers detected within that recording. Let’s represent this set for each recording $R_n$ using $L_n = \{L_{n,m}; m = 1, \ldots, M_n\}$. After diarization of each recording $R_n$, we link the locally unique speakers over all $L_n$ to achieve a set of $U$ globally unique speakers. We store the BW statistics for these globally unique speakers in $G = \{G_{u}; u = 1, \ldots, U\}$. As the global speaker statistics in $G$ have been gathered from across multiple recordings, they can provide the re-diarization of a recording $R_n$ with additional speaker statistics that are not local to that recording. To do this, for each recording $R_n$, we find its set of globally unique speaker statistics that are in $G$. We call this set $D_n$, where:

$$D_n = \{D_{n,m}; m = 1, \ldots, M_n; D_{n,m} \in G\},$$

and $D_n \subseteq G$. We then re-use the baseline annotations to form a new initial segmentation of each recording $R_n$, obtaining $K'_n$ speaker segments $X'_n = \{X'_{n,k}; k = 1, \ldots, K'_n\}$ such that $|X'_n| \geq |L_n|$. We obtain this segmentation by treating the speaker-change points in the baseline annotations as the start of a new segment. As we have the VAD decisions from the previous run, we only need to reapply diarization (Section 3.2) from the clustering stage. To take into account the global information in the re-diarization of a recording $R_n$, we participate the set of speaker statistics in $D_n$ in the clustering of segments in $X'_n$. Hence, rather than scoring and clustering only $K'_n$ segments, we now have $K'_n + M_n$ segments, where $M_n$ is the set of additional speaker statistics in $D_n$ that are relevant to recording $R_n$. We then carry out clustering to obtain a new set of locally unique speakers, ignoring the labels of the added models (as they are not from that recording), obtaining an ideally improved update of the locally unique speaker set $L'_n$. We can now reapply our baseline linking system to the new sets of locally unique speaker statistics $L'_n$ over all $N$ recordings, to achieve attribution. We thus obtain an updated set of globally unique speaker statistics $G' = \{G'_{u}; u = 1, \ldots, U\}$. This completes the re-diarization process, which can iteratively be applied to the updated sets in the same manner. This re-diarization scheme would ideally ensure that the added models serve as initial cluster nodes to which speakers with short utterance samples can be attributed when clustering within a recording. We hypothesise that if ideally pure models are used, this can lower the similarity score between competing segments within a recording by guiding the clustering process.

We applied 10 iterations of recombination to the baseline system. Figure 1 displays the performance of our re-diarization scheme over SAIVT-BNEW in terms of the cumulative change in evaluation metrics at each iteration. This is the change with respect to the baseline metrics, hence the change at iteration 0 is 0.0. In Figure 1, improvements to CP and CC translate to a positive change, while a negative change to the ER metric would indicate improvements to DER and AER, respectively. Figure 1(a) displays the change in CP, CC and DER for diarization, Figure 1(b) reflects this for attribution and Figure 1(c) demonstrates the number of globally unique identities hypothesised by the system.

From Figure 1, our recombination scheme initially improves cluster coverage (CC) for attribution, which in this case translates to an improvement of the AER. This is achieved at a cost of undesirably reducing cluster purity (CP). This trend continues until iteration 3, when our models have become so impure that over-clustering of speakers takes place, thus increasing the error rates. This is because our baseline system is not perfect and provides impure models to the re-diarization stage, which in turn attracts segments spoken by more than one unique speaker. This is reinforced by the lower number of speakers, with respect to the true number of speakers, obtained at each iteration.

4.2. Boosting Cluster Purity by Purification

Our recombination approach cannot provide reliable improvements. This is because the globally unique speakers obtained by the baseline are not pure enough to allow for reliable re-diarization, thus reinforcing errors. To rectify this, we need to present an opportunity for the system to reverse its incorrect clustering decisions. We propose a cluster purification, or purification, scheme. We apply the exact same process as recombination (Section 4.1), however rather than conducting re-diarization of each recording using its associated set of additional models $D_n$, we incorporate the entire global set $G$.

We applied 10 iterations of purification to the baseline system. Figure 2 displays the performance of our purification scheme. It can be seen that purification provides a desirable increase in CP and CC metrics for attribution, after just one iteration. This also translates to a desirable reduction of the error metrics for both diarization and attribution. An instant boost in the number of hypothesised speakers is also observed, bringing this value closer to the true number of speakers in the set. Further application of purification appears to reduce CC, especially for diarization, leading to increased error metrics. This is because once the cluster coverage is decreased, a larger number of speakers will be available that inevitably have shorter utterance samples, which leads to unreliable modeling.

4.3. Joint Purification-Recombination Scheme

We propose merging our purification and recombination schemes to conduct re-diarization. We first apply an iteration of purification, to purify the baseline models. We then counter this outcome by applying an iteration of recombination. We apply 10 iteration of this purification-recombination to the baseline output, with odd iterations representing the outcome of purifi-
5. Discussion and conclusions

We demonstrated that it is possible to reuse a non-ideal annotation of a spoken archive, provided by a real attribution (diarization and linking) system, to boost accuracy across that set.

6. Acknowledgements

This research was supported by an Australian Research Council (ARC) Linkage Grant (No: LP130100110) and the Cooperative Research Centre for Smart Services.

Table 2: JPR diarization and attribution evaluations.

<table>
<thead>
<tr>
<th>Baseline+JPR</th>
<th>ER%</th>
<th>CP%</th>
<th>CC%</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diarization</td>
<td>10.1</td>
<td>84.8</td>
<td>92.6</td>
<td>303</td>
</tr>
<tr>
<td>Attribution</td>
<td>29.7</td>
<td>79.3</td>
<td>78.5</td>
<td>103</td>
</tr>
</tbody>
</table>
7. References


